

**Title:**

Pairwise maximum entropy models for studying large biological systems: when they can and when they can't work

**Abbreviated title:** maximum entropy models for biological systems

**Authors:**

Yasser Roudi,

Gatsby Computational Neuroscience Unit, UCL, UK

Department of Physiology and Biophysics,

Weill Medical College of Cornell University, New York, USA

Sheila Nirenberg,

Department of Physiology and Biophysics,

Weill Medical College of Cornell University, New York, USA

Peter, E. Latham,

Gatsby Computation Neuroscience Unit, UCL, UK

## Abstract

One of the most critical problems we face in the study of biological systems is building accurate statistical descriptions of them. This problem has been particularly challenging because biological systems typically contain large numbers of interacting elements, which precludes the use of standard brute force approaches. Recently, though, several groups have reported that there may be an alternate strategy. The reports show that reliable statistical models can be built without knowledge of all the interactions in a system; instead, pairwise interactions can suffice. These findings, however, are based on the analysis of small subsystems. Here we ask whether the observations will generalize to systems of realistic size, that is, whether pairwise models will provide reliable descriptions of true biological systems. Our results show that, in most cases, they will not. The reason is that there is a crossover in the predictive power of pairwise models: If the size of the subsystem is below the crossover point, then the results have no predictive power for large systems. If the size is above the crossover point, the results do have predictive power. This work thus provides a general framework for determining the extent to which pairwise models can be used to predict the behavior of whole biological systems. Applied to neural data, the size of most systems studied so far is below the crossover point.

# 1 Introduction

Many important questions in biology are fundamentally statistical. For instance, deciphering the neural code requires knowledge of the probability of observing patterns of activity in response to stimuli [1]; determining which features of a protein are important for correct folding requires knowledge of the probability that a particular sequence of amino acids folds naturally [2, 3]; and determining the patterns of foraging of animals and their social and individual behavior requires knowledge of the distribution of food and species over both space and time [4–6].

Building statistical descriptions of biological systems is, however, hard. There are several reasons for this: i) biological systems are composed of large numbers of elements, and so can exhibit a huge number of configurations, in fact, an exponentially large number, ii) the elements typically interact with each other, making it impossible to view the system as a collection of independent entities, and iii) because of technological considerations, the descriptions of biological systems have to be built from very little data. For example, with current technology in neuroscience, we can record simultaneously from only about 100 neurons out of approximately 100 billion in the human brain. So, not only are we faced with the problem of estimating probability distributions in high dimensional spaces, we must make the estimates based on very little information.

Despite these apparent difficulties, recent work has suggested that the situation may be less bleak than it seems. There is evidence that accurate statistical description of systems can be achieved without having to examine all possible configurations [2, 3, 7–11]. One merely has to measure the probability distribution over pairs of elements and use those to build the full distribution. These “pairwise models” potentially offer a fundamental simplification, since the number of pairs is quadratic in the number of elements, not exponential. However, support for the efficacy of pairwise models has, necessarily, come from relatively small subsystems – small enough that the true probability distribution could be measured experimentally, allowing direct comparison of the pairwise distribution to the true one [7–9, 11]. While these studies have provided a key first step, a critical question remains: will the results from the analysis of these small subsystems extrapolate to large ones? That is, if a pairwise model predicts the probability distribution for a subset of the elements in a system, will it also predict the probability distribution for the whole system? Here we find that, for a biologically relevant

class of systems, this question can be answered quantitatively and, importantly, generically – independent of many of the details of the biological system under consideration. And the answer is, generally, “no.” In this paper, we explain, both analytically and with simulations, why this is the case.

## 2 Results

### 2.1 The extrapolation problem

To gain intuition into the extrapolation problem, let us consider a specific example: neuronal spike trains. Figure 1A shows a typical spike train for a small population of neurons. Although the raw spike times provide a complete description, they are not a useful representation, as they are too high-dimensional. Therefore, we divide time into bins and re-represent the spike train as 0s and 1s: 0 if there is no spike in a bin; 1 otherwise (Fig. 1B) [7–9, 11]. For now we assume that the bins are independent (an assumption whose validity we discuss below, and in more detail in Sec. 3). The problem, then, is to find  $p_{true}(\mathbf{r}) \equiv p_{true}(r_1, r_2, \dots, r_N)$  where  $r_i$  is a binary variable indicating no spike ( $r_i = 0$ ) or one or more spikes ( $r_i = 1$ ) on neuron  $i$ . Since this, too, is a high dimensional problem (though less so than the original spike time representation), suppose that we instead construct a pairwise approximation to  $p_{true}$ , which we denote  $p_{pair}$ , for a population of size  $N$ . (The pairwise model derives its name from the fact that it has the same mean and pairwise correlations as the true model.) Our question, then, is: if  $p_{pair}$  is close to  $p_{true}$  for small  $N$ , what can we say about how close the two distributions are for large  $N$ ?

To answer this question quantitatively, we need a measure of distance. The measure we use, denoted  $\Delta_N$ , is defined in Eq. (8) below, but all we need to know about it for now is that if  $\Delta_N = 0$  then  $p_{pair} = p_{true}$ , and if  $\Delta_N$  is near one then  $p_{pair}$  is far from  $p_{true}$ . In terms of  $\Delta_N$ , our main results are as follows: first, for small  $N$ , in what we call the perturbative regime,  $\Delta_N$  is proportional to  $N - 2$ . In other words, as the population size increases, the pairwise model becomes a worse and worse approximation to the true distribution. Second, this behavior is entirely generic: for small  $N$ ,  $\Delta_N$  increases linearly, no matter what the true distribution is. This is illustrated schematically in Fig. 2, which shows the generic behavior of  $\Delta_N$ . The solid red part of the curve is the perturbative regime, where  $\Delta_N$  is a linearly increasing function of  $N$ ; the dashed curves show possible behavior beyond the perturbative regime.

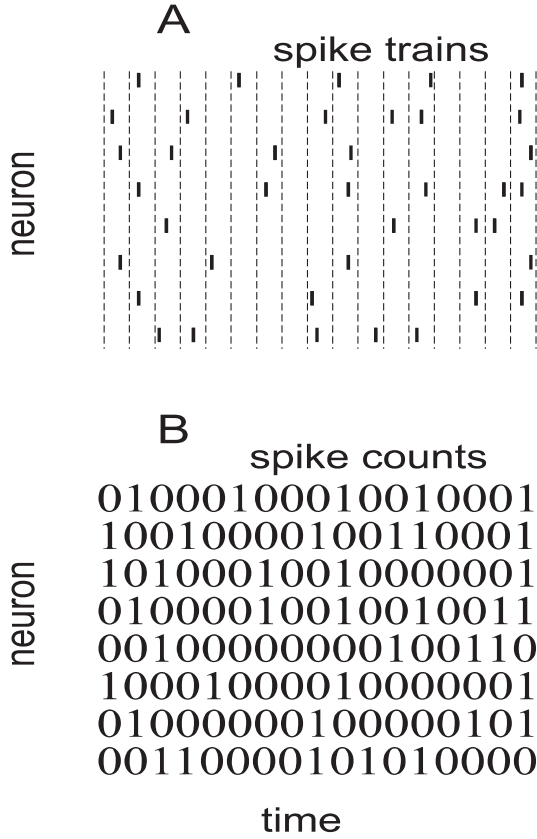


Figure 1: Transforming spike trains to spike count. **A.** Spike rasters. Tick marks indicate spike times; different rows correspond to different neurons. The horizontal dashed lines are the bin boundaries. **B.** Spike count in each bin. In this example the bins are small enough that there is at most one spike per bin, but this is not necessary – one could use bigger bins and have larger spike counts.

These results have an important corollary: if one does an experiment and finds that  $\Delta_N$  is increasing linearly with  $N$ , then one has no information at all about the true distribution. The flip side of this is more encouraging: if one can measure the true distribution for sufficiently large  $N$  that  $\Delta_N$  saturates, as in the dashed blue line in Fig. 2, then one can have some confidence that extrapolation to large  $N$  is meaningful. The implications for the interpretation of experiments is, then, that extrapolation to large  $N$  is valid only if one can analyze data past the perturbative regime.

Under what conditions is a subsystem in the perturbative regime? The answer turns out to be simple: the size of the system,  $N$ , times the average probability of observing a spike in a bin, must be small compared to 1. For example, if the average probability is 1/100, then a system will be in the perturbative regime if the number of neurons is small compared to 100. This last observation would seem to be good news for studies in which spikes are binned across time and temporal correlations are ignored. For such binned spike trains, the probability of a spike can be made arbitrarily small by simply shrinking the time bins, and so the size of

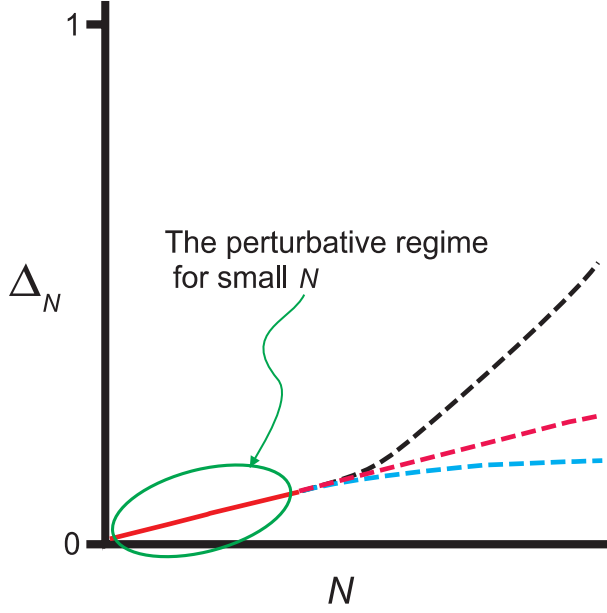


Figure 2: Cartoon illustrating the dependence of  $\Delta_N$  on  $N$ . For small  $N$  there is always a perturbative regime in which  $\Delta_N$  increases linearly with  $N$  (solid red line). When  $N$  becomes large, on the other hand,  $\Delta_N$  may continue increasing with  $N$  (red and black dashed lines) or it may plateau (cyan dashed line), depending on  $p_{true}$ . The observation that  $\Delta_N$  increases linearly with  $N$  does not, therefore, provide much, if any information about the large  $N$  behavior.

the population for which the pairwise model appears good can be made arbitrarily large. The problem with this, though, is that temporal correlations can be ignored only when time bins are large compared to the autocorrelation time. This leads to a kind of catch-22: pairwise models are guaranteed to work well (in the sense that they describe spike trains in which temporal correlations are ignored) if one uses small time bins, but small time bins is the one regime where ignoring temporal correlations is not a valid approximation.

In the next several sections we quantify the qualitative picture presented above: we write down an explicit expression for  $\Delta_N$ , explain why it increases linearly with  $N$ , when  $N$  is small, and provide additional tests, besides assessing the linearity of  $\Delta_N$ , to determine whether or not one is in the perturbative regime.

## 2.2 A measure of goodness of fit

A natural measure of the distance between  $p_{pair}$  and  $p_{true}$  is the Kullback-Leibler (KL) divergence [12], denoted  $D_{KL}(p_{true}||p_{pair})$  and defined as

$$D_{KL}(p_{true}||p_{pair}) = \sum_{\mathbf{r}} p_{true}(\mathbf{r}) \log_2 \frac{p_{true}(\mathbf{r})}{p_{pair}(\mathbf{r})}. \quad (1)$$

The KL divergence is zero if the two distributions are equal; otherwise it is nonzero.

Although the KL divergence is a very natural measure, it is not easy to interpret (except, of course, when it is exactly zero). That is because a nonzero KL divergence tells us that  $p_{pair} \neq p_{true}$ , but it does not give us any real handle on how much we benefit by including the pairwise correlations in our approximation. To make sense of the KL divergence, then, we need something to compare it to. A reasonable reference quantity, used by a number of authors [7–9], is the KL divergence between the true distribution and the independent one, the latter denoted  $p_{ind}$ . The independent distribution, as its name suggests, is a distribution in which the variables are taken to be independent,

$$p_{ind}(r_1, \dots, r_N) = \prod_i p_i(r_i), \quad (2)$$

where  $p_i(r_i)$  is the distribution of the response of the  $i^{\text{th}}$  neuron,  $r_i$ . With this choice for a comparison, we define our measure of goodness of fit as

$$\Delta_N = \frac{D_{KL}(p_{true}||p_{pair})}{D_{KL}(p_{true}||p_{ind})}. \quad (3)$$

Note that the denominator in this expression,  $D_{KL}(p_{true}||p_{ind})$ , is usually referred to as the multi-information [7, 13, 14].

The quantity  $\Delta_N$ , which we introduced in the previous section, lies between 0 and 1, and measures how well a pairwise model does relative to an independent model. If it is 0, the pairwise model is equal to the true model ( $p_{pair}(\mathbf{r}) = p_{true}(\mathbf{r})$ ); if it is near 1, the pairwise model offers little improvement over the independent model; and if it is exactly 1, the pairwise model is equal to the independent model ( $p_{pair}(\mathbf{r}) = p_{ind}(\mathbf{r})$ ), and so offers no improvement. (Our assertion that  $\Delta_N$  cannot exceed 1 assumes that the pairwise model cannot be worse than the independent one, something that is reasonable in practice but not guaranteed in general.)

How do we attach physical meaning to the two divergences  $D_{KL}(p_{true}||p_{pair})$  and  $D_{KL}(p_{true}||p_{ind})$ ? For the latter, we use the fact that, as is easy to show,

$$D_{KL}(p_{true}||p_{ind}) = S_{ind} - S_{true}, \quad (4)$$

where  $S_{ind}$  and  $S_{true}$  are the entropies [15, 16] of  $p_{ind}$  and  $p_{true}$ , respectively, defined, as usual, to be  $S[p] = -\sum_{\mathbf{r}} p(\mathbf{r}) \log_2 p(\mathbf{r})$ . For the former, we use the definition of the KL divergence to write

$$D_{KL}(p_{true}||p_{pair}) = -\sum_{\mathbf{r}} p_{true}(\mathbf{r}) \log_2(p_{pair}(\mathbf{r})) - S_{true} \equiv \tilde{S}_{pair} - S_{true}. \quad (5)$$

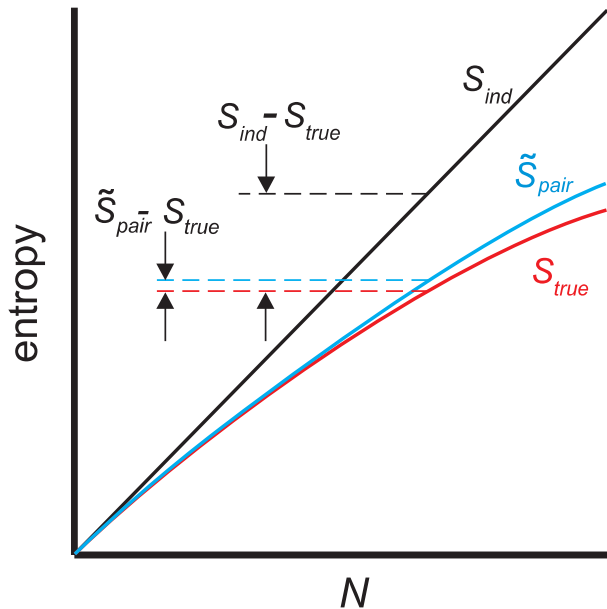


Figure 3: Schematic plot of  $S_{ind}$  (black line),  $\tilde{S}_{pair}$  (cyan line) and  $S_{true}$  (red line). The better the pairwise model, the closer  $\tilde{S}_{pair}$  is to  $S_{true}$ . This is reflected in the cost function  $\Delta_N$ , which is the distance between the red and cyan lines divided by the distance between the red and black lines.

The quantity  $\tilde{S}_{pair}$  has the flavor of an entropy, although it is a true entropy only when  $p_{pair}$  is maximum entropy as well as pairwise (the maximum entropy pairwise model, or maximum entropy model for short; see Eq. (11)). For other pairwise distributions, all we need to know is that  $\tilde{S}_{pair}$  lies between  $S_{true}$  and  $S_{ind}$ , something that is guaranteed by our assumption that the pairwise model is no worse than the independent model.

What Eqs. (4) and (5) tell us is that  $\Delta_N$  (Eq. (3)) is the ratio of the amount of “entropy” *not* explained by the pairwise model to the amount of entropy *not* explained by the independent model. A plot illustrating the relationship between  $\Delta_N$ , the two entropies  $S_{ind}$  and  $S_{true}$ , and the entropy-like quantity  $\tilde{S}_{pair}$ , is shown in Fig. 3.

### 2.3 $\Delta_N$ in the perturbative regime

The extrapolation problem discussed above is the problem of determining  $\Delta_N$  in the large  $N$  limit. This is hard to do in general, but, as we show in Methods, Sec. 5.2, there is a perturbative regime in which it is possible. The small parameter that defines this regime is the average number of spikes per bin; this is written quantitatively as  $N\bar{\nu}\delta t$  where  $\delta t$  is the bin size and  $\bar{\nu}$  the average firing rate. Letting and  $\nu_i$  the firing rate of neuron  $i$ , the latter quantity is given by

$$\bar{\nu} \equiv \frac{1}{N} \sum_i \nu_i. \quad (6)$$

Note that our small parameter depends on  $N$ , which means that for  $\bar{\nu}$  and  $\delta t$  fixed, there is a



maximum population size we can access perturbatively. We return to this point below, as it has major implications for experimental studies.

The first step in the perturbation expansion is to compute the two quantities that make up  $\Delta_N$ :  $D_{KL}(p_{true}||p_{ind})$  and  $D_{KL}(p_{true}||p_{pair})$ . As we show in Methods, Sec. 5.2, these are given by

$$D_{KL}(p_{true}||p_{ind}) = g_{ind}N(N-1) (\bar{\nu}\delta t)^2 + \mathcal{O}((N\bar{\nu}\delta t)^3) \quad (7a)$$

$$D_{KL}(p_{true}||p_{pair}) = g_{pair}N(N-1)(N-2) (\bar{\nu}\delta t)^3 + \mathcal{O}((N\bar{\nu}\delta t)^4) . \quad (7b)$$

The exact forms of the prefactors  $g_{ind}$  and  $g_{pair}$  are given in Eqs. (41) and (43). The details, however, are not so important; the important things to know about them is that they are independent of  $N$  and  $\bar{\nu}\delta t$ , and they depend on the low order statistics of the spike trains:  $g_{ind}$  depends on the second order normalized correlations function, and  $g_{pair}$  depends on the second and third order normalized correlations function, as defined below in Eq. (13). The  $N$ -dependence in the first term on the right hand side of Eq. (7a) has been noted previously [7], although the authors did not compute the prefactor,  $g_{ind}$ .

Inserting Eq. (7) into Eq. (3) (into the definition of  $\Delta_N$ ), we arrive at our main result,

$$\Delta_N = \frac{g_{pair}}{g_{ind}}(N-2) \bar{\nu}\delta t + \mathcal{O}((N\bar{\nu}\delta t)^2) . \quad (8)$$

This expression tells us how  $\Delta_N$  scales with  $N$  in the perturbative regime – the regime in which  $N\bar{\nu}\delta t \ll 1$ . The key observation about this scaling is that it is independent of the details of the true distribution,  $p_{true}$ . This has a very important consequence, one that has major implications for experimental data: if one does an experiment and finds that that  $\Delta_N$  is proportional to  $N-2$ , then the system is in the perturbative regime, and one does not know whether  $p_{pair}$  will remain close to  $p_{true}$  as  $N$  increases. What this means in practical terms is that if one wants to know whether a particular pairwise model is a good one for large systems, it is necessary to consider values of  $N$  that are significantly greater than  $N_c$ , where

$$N_c \equiv \frac{1}{\bar{\nu}\delta t} . \quad (9)$$

We interpret  $N_c$  as the value at which there is a crossover in the behavior of the pairwise model. Specifically, if  $N \ll N_c$ , the system is in the perturbative regime and the pairwise model is not

informative about the large  $N$  behavior, whereas if  $N \gg N_c$ , the system is in a regime in which it may be possible to make inferences about the behavior of the full system.

## 2.4 The dangers of extrapolation

Although the behavior of  $\Delta_N$  in the perturbative regime does not tell us much about its behavior at large  $N$ , it is possible that other quantities that can be calculated in the perturbative regime,  $g_{ind}$ ,  $g_{pair}$ , and  $S_{ind}$  (the last one exactly), are informative, as others have suggested [7]. Here we show that they are also uninformative.

The easiest way to relate the perturbative regime to the large  $N$  regime is to extrapolate Eqs. (7a) and (7b), and ask what their large  $N$  behavior tells us. Generic versions of these extrapolations, plotted on a log-log scale, are shown in Fig. 4A, along with a plot of the independent entropy,  $S_{ind}$  (which is necessarily linear in  $N$ ; see Sec. 5.1). The first thing we notice about the extrapolations is that they do not, technically, have a large  $N$  behavior: one terminates at the point labeled  $N_{ind}$ , which is where  $D_{KL}(p_{true}||p_{ind}) = S_{ind}$  (and thus  $S_{true} = 0$ ; continuing the extrapolation implies negative true entropy); the other at the point labeled  $N_{pair}$ , which is where  $D_{KL}(p_{true}||p_{pair}) = S_{ind}$  (and thus  $S_{true} \leq 0$ , since  $\tilde{S}_{pair} \leq S_{ind}$ ).

Despite the fact that the extrapolations end abruptly, they still might provide information about the large  $N$  regime. For example,  $N_{pair}$  and/or  $N_{ind}$  might be values of  $N$  at which something interesting happens. To see if this is the case, in Fig. 4B we plot the naive extrapolations of  $\tilde{S}_{pair}$  and  $S_{true}$ , as given by Eq. (7), on a linear-linear plot, along with  $S_{ind}$  (solid lines). This plot contains no new information compared to Fig. 4A, but it does elucidate the meaning of the extrapolations. Perhaps its most striking feature is that the naive extrapolation of  $S_{true}$  has a decreasing portion. As is easy to show mathematically, this cannot happen (intuitively, that is because observing one additional neuron cannot decrease the entropy of previously observed neurons). Thus,  $N_{ind}$ , which occurs well beyond the point where the naive extrapolation of  $S_{true}$  is decreasing, has essentially no meaning, something that has been pointed out previously by Bethge et al. [10]. The other potentially important value of  $N$  is  $N_{pair}$ . This, though, suffers from similar problems: either  $N_{pair} > N_{ind}$ , in which case the entropy is negative, or it crosses the green curve in Fig. 4A from below, meaning  $\Delta_N > 1$ . Either way, it also doesn't have much meaning.

How do the naively extrapolated entropies – the solid lines in Fig. 4B – compare to the

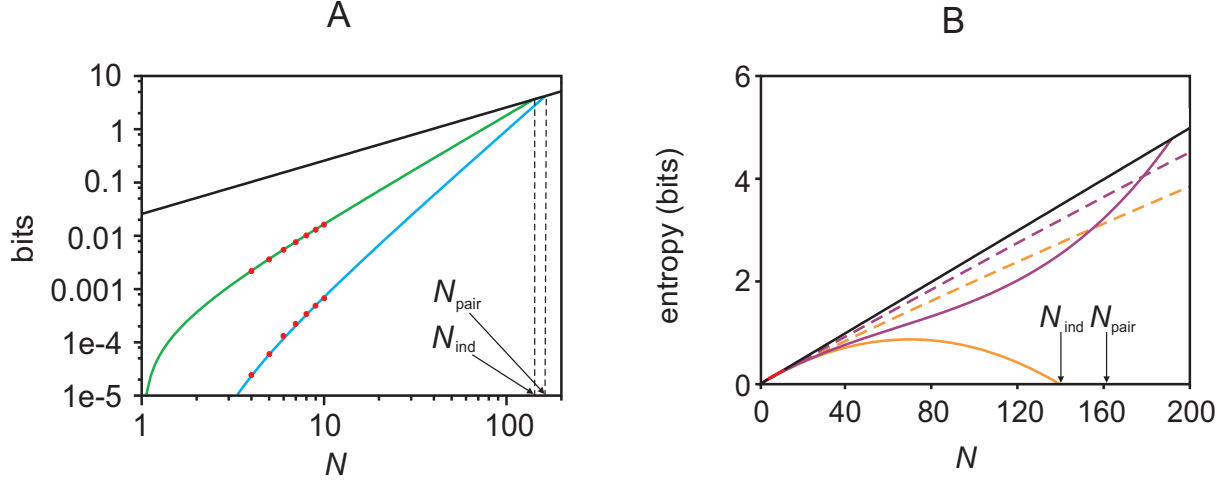


Figure 4: Cartoon showing the extrapolation of entropy and KL divergences, and illustrating why the two natural quantities derived from it,  $N_{ind}$  and  $N_{pair}$ , occur beyond the point at which the extrapolation is meaningful. **A.** Extrapolations on a log-log scale. Black:  $S_{ind}$  versus  $N$ ; green: extrapolation of  $D_{KL}(p_{true}||p_{ind})$ ; cyan: extrapolation of  $D_{KL}(p_{true}||p_{maxent})$ . The red points are the data. The points  $N_{ind}$  and  $N_{pair}$  label the intersections of the two extrapolations with the independent entropy,  $S_{ind}$ . **B.** Extrapolation of the entropies rather than the KL divergences, plotted on a linear-linear scale. The data, again shown in red, is barely visible in the lower left hand corner. Black:  $S_{ind}$  versus  $N$ ; solid maroon: extrapolation of  $\tilde{S}_{pair}$ ; solid orange: extrapolation of  $S_{true}$ . The dashed maroon and orange lines are the extrapolations of the true pairwise “entropy” and true entropy, respectively.

true entropies? To answer this, in Fig. 4B we show the true behavior of  $S_{true}$  and  $\tilde{S}_{pair}$  versus  $N$  (dashed lines). Note that  $S_{true}$  is asymptotically linear in  $N$ , even though the neurons are correlated, a fact that forces  $\tilde{S}_{pair}$  to be linear in  $N$ , as it is sandwiched between  $S_{true}$  and  $S_{ind}$ . (The asymptotically linear behavior of  $S_{true}$  is typical, even in highly correlated systems. Although this is not always appreciated, it is easy to show; see Sec. 5.1.) Comparing the dashed and solid lines, we see that the naively extrapolated and true entropies, and thus the naively extrapolated and true values of  $\Delta_N$ , have extremely different behavior. This further suggests that there is very little connection between the perturbative and large  $N$  regimes.

These observations can be summarized by noting that  $g_{ind}$  and  $g_{pair}$  depend only on correlation coefficients up to third order (see Eqs. (41) and (43)), whereas the large  $N$  behavior depends on correlations at all orders. Thus, since  $g_{ind}$  and  $g_{pair}$  tell us very little, if anything, about higher order correlations, it is not surprising that they tell us very little about the behavior of  $\Delta_N$  in the large  $N$  limit.

## 2.5 Numerical simulations

To check that our perturbation expansions, Eqs. (7) and (8), are correct, and to investigate the regime in which they are valid, we performed numerical simulations. We generated, from synthetic data, a set of true distributions, computed  $D_{KL}(p_{true}||p_{ind})$ ,  $D_{KL}(p_{true}||p_{pair})$ , and  $\Delta_N$  numerically for each of them, and compared to the values predicted by Eqs. (7) and (8). The results are shown in Fig. 5. Before discussing that figure, though, we explain our procedure for constructing true distributions.

The set of true distributions we used were generated from a third order model (so named because it includes up to third order interactions). This model has the form

$$p_{true}(r_1, \dots, r_{N^*}) = \frac{1}{Z_{true}} \exp \left[ \sum_i h_i^{true} r_i + \sum_{i < j} J_{ij}^{true} r_i r_j + \sum_{i < j < k} K_{ijk}^{true} r_i r_j r_k \right] \quad (10)$$

where  $Z_{true}$  is a normalization constant, chosen to ensure the the probability distribution sums to 1, and the sums over  $i$ ,  $j$  and  $k$  run from 1 to  $N^*$ . The parameters  $h_i^{true}$ ,  $J_{ij}^{true}$  and  $K_{ijk}^{true}$  were chosen by sampling from distributions (see Methods, Sec. 5.3), which allowed us to generate many different true distributions.

For a particular simulation (corresponding to a column in Fig. 5), we generated a true distribution with  $N^* = 15$ , randomly chose 5 neurons, and marginalized over them. This gave us

a 10-neuron true distribution. True distributions with  $N < 10$  were constructed by marginalizing over additional neurons within our 10-neuron population. To achieve a representative sample, we considered all possible marginalizations (of which there are  $10 \text{ choose } N$ , or  $10!/[N!(10 - N)!]$ ). The results in Fig. 5 are averages over these marginalizations.

For neural data, the most commonly used pairwise model is the maximum entropy model. Therefore, we use that one here. To emphasize the maximum entropy nature of this model, we replace the label “*pair*” that we have been using so far with “*maxent*.” The maximum entropy distribution has the form

$$p_{maxent}(\mathbf{r}) = \frac{1}{Z} \exp \left[ \sum_i h_i r_i + \sum_{i < j} J_{ij} r_i r_j \right]. \quad (11)$$

Fitting this distribution requires that we choose the  $h_i$  and  $J_{ij}$  so that the first and second moments match those of the true distribution. Quantitatively, these conditions are

$$\langle r_i \rangle_{maxent} = \langle r_i \rangle_{true} \quad (12a)$$

$$\langle r_i r_j \rangle_{maxent} = \langle r_i r_j \rangle_{true} \quad (12b)$$

where the angle brackets,  $\langle \dots \rangle_{maxent}$  and  $\langle \dots \rangle_{true}$ , represent averages with respect to  $p_{maxent}$  and  $p_{true}$ , respectively. Once we have  $h_i$  and  $J_{ij}$  that satisfy Eq. (12), we calculate the KL divergences, Eqs. (7), and use those to compute  $\Delta_N$ .

The results are shown in Fig. 5. The rows correspond to our three quantities of interest:  $D_{KL}(p_{true}||p_{ind})$ ,  $D_{KL}(p_{true}||p_{pair})$ , and  $\Delta_N$  (top to bottom). The columns correspond to different values of  $\overline{\nu}\delta t$ , with the smallest  $\overline{\nu}\delta t$  on the left and the largest on the right. Red circles are the actual values of these quantities; blue ones are the predictions from Eqs. (7) and (8).

As suggested by our perturbation analysis, the smaller the value of  $\overline{\nu}\delta t$ , the better the agreement between computed and predicted values. Our simulations corroborate this: the left column of Fig. 5 has  $\overline{\nu}\delta t = 0.024$ , and agreement is almost perfect out to  $N = 10$ ; the middle column has  $\overline{\nu}\delta t = 0.029$ , and agreement is almost perfect out to  $N = 7$ ; and the right column has  $\overline{\nu}\delta t = 0.037$ , and agreement is not good for any value of  $N$ .

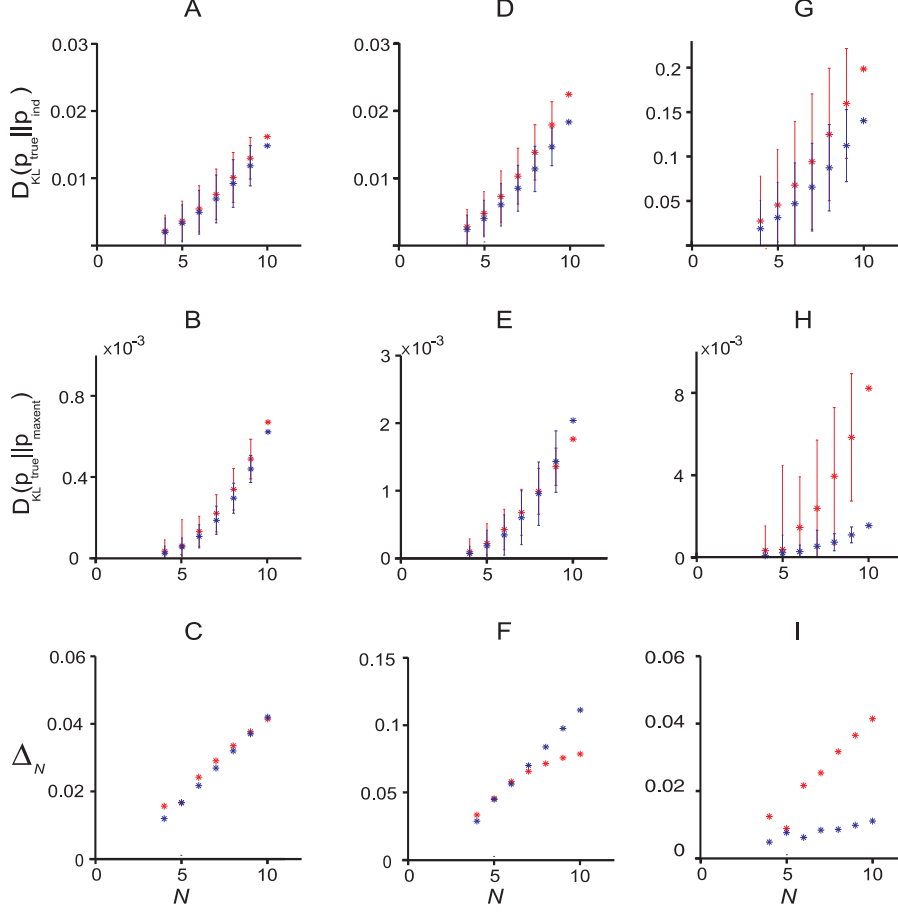


Figure 5: The  $N$  dependence of the KL divergences and the goodness of fit,  $\Delta_N$ . Data was generated from a third order model, as explained in Sec. 5.3 (Methods), and fit to maximum entropy pairwise model and independent models. All data points correspond to averages over marginalizations of the true distribution (see text for details). The red points were computed directly from the model fits, Eqs. (1), (3) and (4); the blue points are predictions of the perturbative expansions, Eqs. (7) and (8). The three columns correspond to  $\bar{\nu}\delta t = 0.024$ ,  $0.029$ , and  $0.037$ , from left to right. **A, B, C** ( $\bar{\nu}\delta t = 0.024$ ). Predictions from the perturbative expansion are in good agreement with the measurements up to  $N = 10$ , indicating that the data is in the perturbative regime. **D, E, F** ( $\bar{\nu}\delta t = 0.029$ ). Predictions from the perturbative expansion are in good agreement with the measurements up to  $N = 7$ , indicating that the data is only partially in the perturbative regime. **G, H, I** ( $\bar{\nu}\delta t = 0.037$ ). Predictions from the perturbative expansion are not in good agreement with the measurements, even for small  $N$ , indicating that the data is outside the perturbative regime.

These results validate the perturbation expansions in Eqs. (7) and (8), and show that those expansions provide sensible predictions – at least for some parameters. They also suggest a natural way to assess the significance of one’s data: plot  $D_{KL}(p_{true}||p_{ind})$ ,  $D_{KL}(p_{true}||p_{pair})$ , and  $\Delta_N$  versus  $N$ , and look for agreement with the predictions of the perturbation expansion. If agreement is good, as in the left column of Fig. 5, then one is in the perturbative regime, and one knows very little about the true distribution. If, on the other hand, agreement is bad, as in the right column, then one is out of the perturbative regime, and there is hope of extracting meaningful information about the relationship between the true and pairwise models.

That said, the qualifier “at least for some parameters” is an important one. This is because the perturbation expansion is essentially an expansion that depends on the normalized correlation coefficients, and there is an underlying assumption that they don’t exhibit pathological behavior. The  $k^{\text{th}}$  normalized correlation coefficient for the distribution  $p$ , denoted  $\rho_{i_1 i_2 \dots i_k}^p$ , is written

$$\rho_{i_1 i_2 \dots i_k}^p \equiv \frac{\langle (\langle r_{i_1} - \langle r_{i_1} \rangle_p ) (\langle r_{i_2} - \langle r_{i_2} \rangle_p ) \dots (\langle r_{i_k} - \langle r_{i_k} \rangle_p ) \rangle_p}{\langle r_{i_1} \rangle_p \langle r_{i_2} \rangle_p \dots \langle r_{i_k} \rangle_p}. \quad (13)$$

A potentially problematic feature of the correlation coefficients is that the denominator is a product over mean activities. If the mean activities are small, the denominator can become very small, leading to very large correlation coefficients. Although our perturbation expansion is always valid for sufficiently small time bins (because the correlation coefficients eventually becomes independent of bin size; see Methods, Sec. 5.4), “sufficiently small” can depend in detail on the parameters. For instance, at the maximum population size tested ( $N = 10$ ) and for the true distributions that had  $\overline{v}\delta t < 0.03$ , the absolute error of the prediction had a median of approximately 16%. However, about 11% of the runs had errors larger than 60%. Thus, the exact size of the small parameter at which the perturbative expansion breaks down can depend on the details of the true distribution.

## 2.6 Local fields and pairwise couplings have a simple dependence on firing rates and correlation coefficients in the perturbative regime

Estimation of the KL divergences and  $\Delta_N$  from real data can be hard, in the sense that it takes a large amount of data for them to converge to their true values. We therefore provide a second

set of relationships that can be used to determine whether or not a particular data set is in the perturbative regime. These relationships are between the parameters of the maximum entropy model, the  $h_i$  and  $J_{ij}$ , and the mean activity and normalized second order correlation coefficient (the latter defined in Eq. (13)).

Since the quantity  $\overline{\nu}\delta t$  plays a central role in our analysis, we replace it with a single parameter, which we denote  $\delta$ ,

$$\delta \equiv \overline{\nu}\delta t. \quad (14)$$

In terms of this parameter, we find, (using the same perturbative approach that led us to Eqs. (7) and (8); see Sec. 5.2.2), that

$$h_i = \log [\langle r_i \rangle^{-1} - 1] + \mathcal{O}(N\delta) \quad (15a)$$

$$J_{ij} = \log [1 + \rho_{ij}] + \mathcal{O}(N\delta) \quad (15b)$$

where  $\rho_{ij}$ , the normalized second order correlation coefficient, is defined in Eq. (13) with  $k = 2$ ; it is given explicitly by

$$\rho_{ij} = \frac{\langle r_i r_j \rangle - \langle r_i \rangle \langle r_j \rangle}{\langle r_i \rangle \langle r_j \rangle}. \quad (16)$$

(We don't need a superscript on  $\rho$  or a subscript on the angle brackets because the first and second moments are the same under the true and pairwise distributions.)

Equation (15) tells us that the  $N$ -dependence of the  $h_i$  and  $J_{ij}$ , the local fields and pairwise couplings, are very weak. In Fig. 6 we confirm these theoretical predictions through numerical simulations.

As an aside, we should point out that the  $N$ -dependence is a function of the variables used to represent the firing patterns. Here we use 0 for silence and 1 for firing, but another, possibly more common, representation, derived from the Ising model and used in a number of studies [7, 9, 11], is to use  $-1$  for silence and  $+1$  for firing. This amounts to making the change of variables  $s_i = 2r_i - 1$ . In terms of  $s_i$ , the maximum entropy model has the form  $p(\mathbf{r}) \sim \exp [\sum_i h_i^{ising} s_i + \sum_{i < j} J_{ij}^{ising} s_i s_j]$  where  $h_i^{ising}$  and  $J_{ij}^{ising}$  are given by

$$h_i^{ising} = \frac{h_i}{2} + \sum_{j \neq i} \frac{J_{ij}}{4} \quad (17a)$$

$$J_{ij}^{ising} = \frac{J_{ij}}{4}. \quad (17b)$$



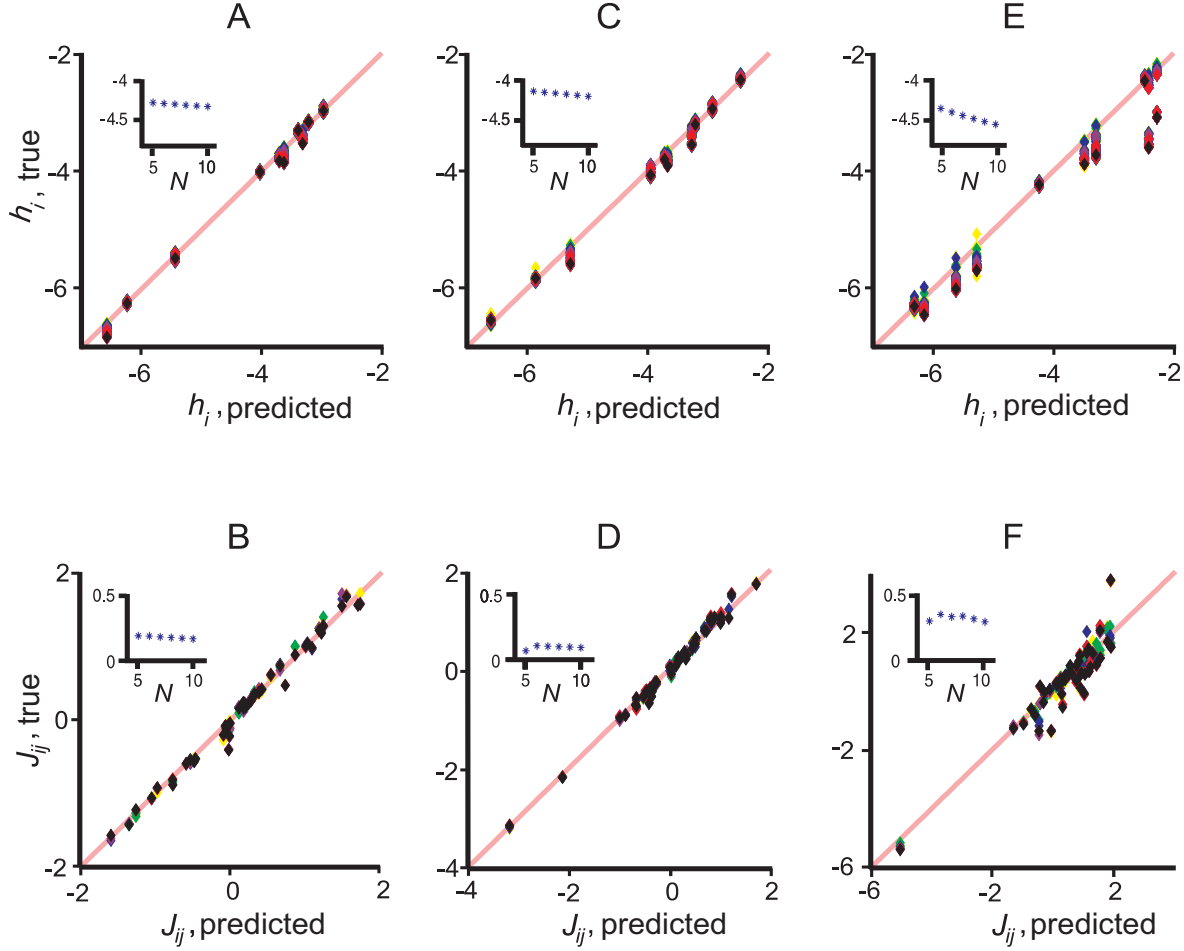


Figure 6: Comparison of the true local fields ( $h_i$ , top row) and pairwise interactions ( $J_{ij}$ , bottom row), to the predictions from the perturbation expansion, Eq. (15). Values of  $N$  ranging from 5 to 10 are shown, with different colors corresponding to different  $N$ s. For each value of  $N$ , fits are shown for 45 realization of the true distribution. Insets show the  $N$ -dependence of the mean local fields (top) and mean pairwise interactions (bottom). The three columns correspond exactly to the columns in Fig. 5. **A, B** ( $\bar{\nu}\delta t = 0.024$ ). There is a very good match between the theoretical and fit values of both local fields and pairwise interactions. **C, D** ( $\bar{\nu}\delta t = 0.029$ ). Even though  $\bar{\nu}\delta t$  has increased, the match is still good. **E, F** ( $\bar{\nu}\delta t = 0.037$ ). The predicted and fit local fields and pairwise interactions do not match as well as the cases shown in A,B, C and D. There is also now a strong  $N$ -dependence in the mean local fields, and a somewhat weaker dependence in the pairwise interactions. This indicates that the perturbative expansion is breaking down.

The second term on the right hand side of Eq. (17a) is proportional to  $N - 1$ , which means the local fields in the Ising representation acquire a linear  $N$ -dependence that was not present in our 0/1 representation. The two studies that reported the  $N$ -dependence of the local fields [7, 9] used this representation, and, as predicted, their local fields had a component that was linear in  $N$ .

Equation (15b) does more than just predict a lack of  $N$ -dependence; it also provides a functional relationship between the pairwise couplings and the normalized pairwise correlations function,  $\rho_{ij}$ . In Figs. 7A-C we plot the pairwise couplings,  $J_{ij}$ , versus the normalized pairwise correlation coefficient,  $\rho_{ij}$  (blue dots), along with the prediction from Eq. (15b) (black line). Consistent with our predictions, the data in Figs. 7A-C essentially follows a line – the one predicted by Eq. (15b).

A relationship between the pairwise couplings and the correlations coefficients has been sought previously, but for the more standard Pearson correlation coefficient [7, 9, 11]. Our analysis explains why it was not found. The Pearson correlation coefficient, denoted  $c_{ij}$ , is given by

$$c_{ij} \equiv \frac{\langle r_i r_j \rangle - \langle r_i \rangle \langle r_j \rangle}{[(\langle r_i^2 \rangle - \langle r_i \rangle^2)(\langle r_j^2 \rangle - \langle r_j \rangle^2)]^{1/2}}. \quad (18)$$

In the small  $\langle r_i \rangle$  limit – the limit of interest – the right hand side of Eq. (18) is approximately equal to  $[\langle r_i \rangle \langle r_j \rangle]^{1/2} \rho_{ij}$ . Because  $[\langle r_i \rangle \langle r_j \rangle]^{1/2}$  depends on the local fields,  $h_i$  and  $h_j$  (see Eq. (15a)) *and* there is a one-to-one relationship between  $\rho_{ij}$  and  $J_{ij}$  (Eq. (15b)), there can't be a one-to-one relationship between  $c_{ij}$  and  $J_{ij}$ . We verify the lack of a relationship in Figs. 7D-E, where we again plot  $J_{ij}$ , but this time versus the standard correlation coefficient,  $c_{ij}$ . As predicted, the data in Figs. 7D-E is scattered over two dimensions. This suggests that  $\rho_{ij}$ , not  $c_{ij}$ , is the natural measure of the correlation between two neurons when they have a binary representation, something that has also been suggested by Amari based on information-geometric arguments [17].

Note that the lack of a simple relationship between the pairwise couplings and the standard correlation coefficient has been a major motivation in building maximum entropy models [7, 11]. This is for good reason: if there is a simple relationship, knowing the  $J_{ij}$ 's adds essentially

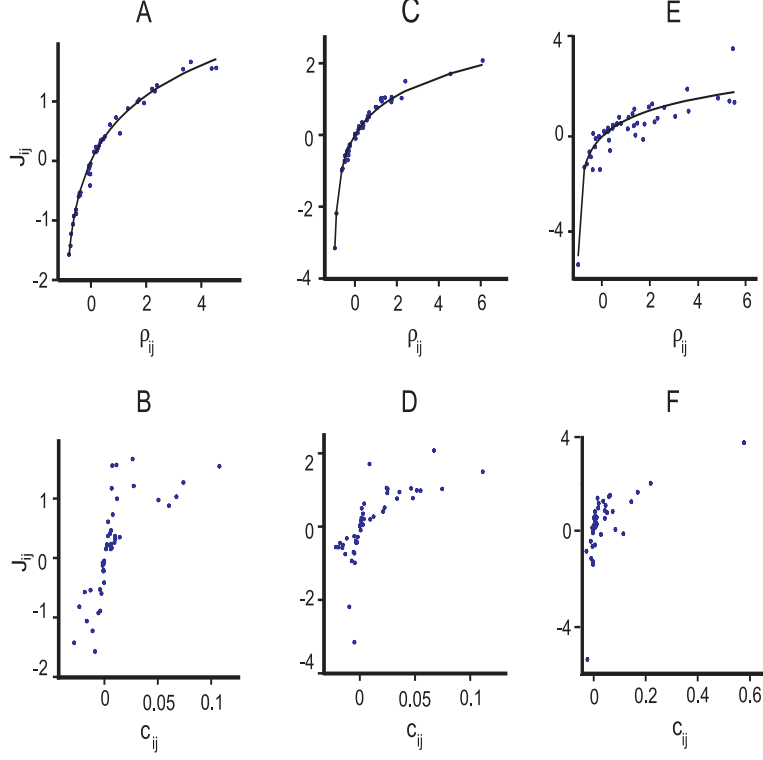


Figure 7: Pairwise couplings versus pairwise correlations, showing that there is a simple relation between  $J_{ij}$  and  $\rho_{ij}$  but not between  $J_{ij}$  and  $c_{ij}$ . Top row:  $J_{ij}$  versus the normalized coefficients,  $\rho_{ij}$  (blue points), along with predicted relationship, via Eq. (15b) (black line). Bottom row:  $J_{ij}$  versus the Pearson correlation coefficients,  $c_{ij}$ , Eq. (18) (blue points). The three columns correspond exactly to the columns in Fig. 5, for which  $\tau\delta t = 0.024, 0.029$ , and  $0.037$ , from left to right. The prediction in the top row (black line) matches the data well, even in the rightmost column.

nothing. Thus, plotting  $J_{ij}$  versus  $\rho_{ij}$  (but not  $c_{ij}$ ) is an important test of one’s data, and if the two quantities fall on the curve predicted by Eq. (15b), the maximum entropy model is adding very little information, if any.

### 3 Is there anything wrong with using small time bins?

An outcome of our perturbative approach is that the goodness of fit measure,  $\Delta_N$ , decreases linearly with bin size (see Eq. (3)). This suggests that one could make the pairwise model look better and better simply by making the bin size smaller and smaller. Is there anything wrong with this? The answer is yes, for reasons we discussed in Sec. 2.1; here we emphasize and expand on this issue, as it is an important one for making sense of experimental results.

The problem arises because what we have been calling the “true” distribution is not really the true distribution of spike trains. It is the distribution assuming independent time bins, an assumption that becomes worse and worse as we make the bins smaller and smaller. (We use this potentially confusing nomenclature primarily because all studies of neuronal data carried out so far have assumed temporal independence, and compared the pairwise distribution to the temporally independent – but still correlated across neurons – distribution [7–9, 11]. In addition, the correct name “true under the assumption of temporal independence,” is unwieldy.) Here we quantify how much worse. In particular, we show that if one uses time bins that are small compared to the characteristic correlation time in the spike trains, the pairwise model will not provide a good description of the data. Essentially, we show that, when the time bins are too small, the error one makes in ignoring temporal correlations is larger than the error one makes in ignoring correlations across neurons.

As usual, we divide time into bins of size  $\delta t$ . However, because we are dropping the independence assumption, we use  $\mathbf{r}^t$ , rather than  $\mathbf{r}$ , to denote the response in bin  $t$ . The full probability distribution over all time bins is denoted  $\mathcal{P}(\mathbf{r}^1, \dots, \mathbf{r}^M)$ . Here  $M$  is the number of bins; it is equal to  $T/\delta t$  for spike trains of length  $T$ .

If time bins are approximately independent, we can write

$$\mathcal{P}(\mathbf{r}^1, \dots, \mathbf{r}^M) \approx \prod_t p_{true}(\mathbf{r}^t), \quad (19)$$

and if the pairwise model is a good one, we have

$$p_{true}(\mathbf{r}^t) \approx p_{pair}(\mathbf{r}^t). \quad (20)$$

Combining Eqs. (19) and Eq. (20) then gives us an especially simple expression for the full probability distribution:  $\mathcal{P}(\mathbf{r}^1, \dots, \mathbf{r}^M) \approx \prod_t p_{pair}(\mathbf{r}^t)$ .

The problem with small time bins lies in Eq. (19): the right hand side is a good approximation to the true distribution when the time bins are large compared to the spike train correlation time, but it is a bad approximation when the time bins are small (because adjacent time bins become highly correlated). To quantify how bad, we compare the error one makes assuming independence across time to the error one makes assuming independence across neurons. The ratio of those two errors, denoted  $\gamma$ , is given by

$$\gamma = \frac{D_{KL}(\mathcal{P}(\mathbf{r}^1, \dots, \mathbf{r}^M) || \prod_t p_{pair}(\mathbf{r}^t))}{MD_{KL}(p(\mathbf{r}) || p_{ind}(\mathbf{r}))}. \quad (21)$$

It is relatively easy to compute  $\gamma$  in the limit of small time bins (see Sec. 5.5), and we find that

$$\gamma = \Delta_N + (M - 1) + \frac{\log_2 M}{g_{pair}\delta}. \quad (22)$$

As expected, this reduces to our old result,  $\Delta_N$ , when there is only one time bin ( $M = 1$ ). When  $M$  is larger than 1, however (which is, of course, the case of interest), the second term is always at least one, and for small bin size, the third term is much larger than one. Consequently, if we use bins that are small compared to the temporal correlation time of the spike train, the pairwise model will do a very bad job describing the full, temporally correlated spike trains.

## 4 Discussion

Probability distributions over the configurations of biological systems are extremely important quantities. However, because of the large number of interacting elements comprising such systems, these distributions can almost never be determined directly from experimental data. Using parametric models to approximate the true distribution is the only existing alternative. While such models are promising, they are typically applied only to small subsystems, not the full system. This raises the question: are they good models of the full system?.

We answered this question for a class of parametric models known as pairwise models. We focused on a particular application, neuronal spike trains, and our main result is as follows: if

one were to record spikes from multiple neurons, use sufficiently small time bins and a sufficiently small number of cells, and assume temporal independence, then a pairwise model would almost always succeed in matching the true (but temporally independent) distribution – whether or not it would match the true (but still temporally independent) distribution for large time bins or a large number of cells. In other words, pairwise models in the “sufficiently small” regime, what we refer to as the perturbative regime, have almost no predictive value for what will happen with large populations. This makes extrapolation from small to large systems dangerous.

This observation is important because pairwise models, and in particular maximum entropy pairwise models, have recently attracted a great deal of attention: they have been applied to salamander and guinea pig retinas [7], primate retina [8], primate cortex [9], cultured cortical networks [7], and cat visual cortex [11]. These studies have mainly operated close to the perturbative regime. For example, Schneidman et al. [7] had  $N\overline{\nu}\delta t \approx 0.35$ , Tang et al. [9] had  $N\overline{\nu}\delta t \approx 0.06$  to  $0.4$  (depending on the preparation), and Yu et al. had  $N\overline{\nu}\delta t \approx 0.2$ . For these studies, then, it would be hard to justify extrapolating to large populations.

The study by Shlens et al. [8], on the other hand, might be more amenable to extrapolation. This is because spatially localized stimuli were used to stimulate retinal ganglion cells, for which a nearest neighbor maximum entropy models provided a good fit to their data. (Nearest neighbor means  $J_{ij}$  is zero unless cell  $i$  and cell  $j$  are adjacent.) As is not hard to show, for nearest neighbor models the small parameter in the perturbative expansion is  $K\overline{\nu}\delta t$  where  $K$  is the number of nearest neighbors. Since  $K$  is fixed, independent of the population size, the small parameter will not change as the population size increases. Thus, Shlens et al. may have tapped into the large population behavior even though they considered only a few cells at a time in their analysis.

#### 4.1 Time bins and population size

That the pairwise model is always good if  $N\overline{\nu}\delta t$  is sufficiently small has strong implications: if we want to build a good model for a particular  $N$ , we can simply choose a bin size that is small compared to  $1/N\overline{\nu}$ . However, one of the assumptions in all pairwise models used on neural data is that bins at different times are independent. This produces a tension between small time bins and temporal independence: small time bins essentially ensure that a pairwise model will provide a close approximation to a model with independent bins, but they make adjacent bins

highly correlated. Large time bins come with no such assurance, but they make adjacent bins independent. Unfortunately, this tension is often unresolvable in large populations, in the sense that pairwise models are assured to work only up to populations of size  $1/(\bar{\nu}\tau_{\text{corr}})$  where  $\tau_{\text{corr}}$  is the typical correlation time. Given that  $\bar{\nu}$  is at least several Hz, for experimental paradigms in which the correlation time is more than a few hundred ms,  $1/(\bar{\nu}\tau_{\text{corr}})$  is about one, and this assurance does not apply to even moderately sized populations of neurons.

These observations are especially relevant for studies that use small time bins to model spike trains driven by natural stimuli. This is because the long correlation times inherent in natural stimuli are passed on to the spike trains, so the assumption of independence across time (which is required for the independence assumption to be valid) breaks badly. Knowing that these models are successful in describing spike trains under the independence assumption, then, does not tell us whether they will be successful in describing full, temporally correlated, spike trains. Note that for studies that use stimuli with short correlation times (e.g., non-natural stimuli such as white noise), the temporal correlations in the spike trains are likely to be short, and using small time bins may be perfectly valid.

The only study that has investigated the issue of temporal correlations in maximum entropy models does indeed support the above picture [9]: for the parameters used in that study ( $N\bar{\nu}\delta t = 0.06$  to  $0.4$ ), the maximum entropy pairwise model provided a good fit to the data ( $\Delta_N$  was typically smaller than  $0.1$ ), but it did not do a good job modeling the temporal structure of the spike trains.

## 4.2 Other biological problems that have been approached with pairwise models, e.g, protein folding

As mentioned in the Introduction, in addition to the studies on neuronal data, studies on protein folding have also emphasized the role of pairwise interactions [2, 3]. Briefly, proteins consist of strings of amino acids, and a major question in structural biology is: what is the probability distribution of amino acid strings in naturally folding proteins? One way to answer this is to approximate the full probability distribution of naturally folding proteins from knowledge of single-site and pairwise distributions. One can show that there is a perturbative regime for proteins as well. This can be readily seen using the celebrated HP protein model [18]), where a protein is composed of only two types of amino acids: if, at each site, one amino acid type is

preferred and occurs with high probability, say  $1 - \delta$  ( $\delta \ll 1$ ), then a protein of length shorter than  $1/\delta$  will be in the perturbative regime, and, therefore, a good match between the true distribution and the pairwise distribution for such a protein is virtually guaranteed.

Fortunately, though, the properties of real proteins generally prevent this from happening: at the majority of sites in a protein, the distribution of amino acids is *not* sharply peaked around one amino acid. Even for those sites that are sharply peaked (the evolutionarily-conserved sites), the probability of the most likely amino acid rarely exceeds 90% [19, 20]. This puts proteins consisting of only a few amino acids out of the perturbative regime, and puts longer proteins – the ones usually studied using pairwise models – well out of it.

This difference is fundamental: because many of the studies that have been carried out on neural data were in the perturbative regime, the conclusions of those studies – specifically, the conclusion that pairwise models provide accurate descriptions of large populations of neurons – is not yet supported. This is not the case for the protein studies, because they are not in the perturbative regime. Thus, the evidence that pairwise models provide accurate descriptions of protein folding remain strong and exceedingly promising.

### 4.3 Outlook

We have developed a framework for assessing the validity of pairwise models applied to small systems. Essentially, we developed a set of tests to determine whether one’s data is in the perturbative regime, a regime in which extrapolation to large populations is not warranted. This should serve as a useful guide, not just for analyzing experiments, but also for designing them.

Although our framework is general, we focused primarily on its application to neural data. One of our main results is that the bin size carries an important tradeoff: if the bin size is small, then pairwise models work well, but at the price of ignoring temporal correlations; if the bin size is large enough so that adjacent bins are weakly correlated, then there is no guarantee that pairwise models will work at all. Pairwise models with small time bins, therefore, might be rescued by a small modification: take into account correlations across time as well as neurons. This would increase the complexity of the models, but the amount of data one needs to fit them would still not be so large, as only pairwise correlations and single neuron firing rates need to be estimated. Whether this modification would produce good models is not clear, but if it did



it would bring us much closer to a fundamental understanding of neural systems.

## 5 Methods

### 5.1 The behavior of the true entropy in the large $N$ limit

To understand how the true entropy behaves in the large  $N$  limit, we need only express the difference of the entropies as a mutual information. Using  $S_N$  to denote the true entropy of  $N$  neurons and  $I(1; N)$  to denote the mutual information between one neuron and the other  $N$  neurons in a population of size  $N + 1$ , we have

$$(S_N + S_1) - S_{N+1} = I(1; N) \quad \Rightarrow \quad S_{N+1} - S_N = S_1 - I(1; N). \quad (23)$$

If knowing the activity of  $N$  neurons does not constrain the firing of neuron  $N + 1$ , then the single neuron entropy,  $S_1$ , will exceed the mutual information,  $I(1; N)$ , and the entropy will be an increasing function of  $N$ . For the entropy to be linear in  $N$ , all we need is that the mutual information saturates with  $N$ . Because of synaptic failures, this is a reasonable assumption for networks of neurons: even if we observed all the neurons, there is still residual noise associated with uncertainty about which vesicles release neurotransmitter. Thus, using  $I(1; \infty)$  to denote the asymptotic value of the mutual information and  $\langle S_1 \rangle$  to denote the average single-neuron entropy, we have

$$S_N = N[\langle S_1 \rangle - I(1; \infty)] + \text{corrections}, \quad (24)$$

where the corrections are sublinear in  $N$ .

### 5.2 Perturbative Expansion

Our main quantitative result, given in Eq. (7), is that the KL divergence between the true distribution and both the independent and pairwise distributions can be computed perturbatively as an expansion in powers of  $N\delta$ . Here we carry out this expansion, and derive explicit expressions for the quantities  $g_{ind}$  and  $g_{pair}$ .

To simplify our notation, here we use  $p(\mathbf{r})$  for the true distribution. The critical step in computing the KL divergences perturbatively is to use the Sarmanov-Lancaster expansion [21–

25] for  $p(\mathbf{r})$ ,

$$p(\mathbf{r}) = p_{ind}(\mathbf{r}) (1 + \xi_p(\mathbf{r})) \quad (25)$$

where

$$p_{ind}(\mathbf{r}) = \frac{\exp \sum_i \mathcal{H}_i^p r_i}{\prod_i [1 + \exp(\mathcal{H}_i^p r_i)]} \quad (26a)$$

$$\xi_p(\mathbf{r}) \equiv \sum_{i < j} \mathcal{J}_{ij}^p \delta r_i \delta r_j + \sum_{i < j < k} \mathcal{K}_{ijk}^p \delta r_i \delta r_j \delta r_k + \dots \quad (26b)$$

$$\delta r_i \equiv r_i - \bar{r}_i \quad (26c)$$

$$\bar{r}_i \equiv (1 + \exp(-\mathcal{H}_i))^{\perp} . \quad (26d)$$

This expansion has a number of important, but not immediately obvious, properties. First, as can be shown by a direct calculation,

$$\langle r_i \rangle_p = \langle r_i \rangle_{ind} = \bar{r}_i \quad (27)$$

where the subscripts  $p$  and  $ind$  indicate an average with respect to  $p(\mathbf{r})$  and  $p_{ind}(\mathbf{r})$ , respectively. This has an immediate corollary,

$$\langle \delta r_i \rangle_{ind} = 0 . \quad (28)$$

This last relationship is important, because it tells us that if a product of  $\delta r$ 's contains any terms linear in one of the  $\delta r_i$ , the whole product averages to zero under the independent distribution. This can be used to show that

$$\langle \xi_p(\mathbf{r}) \rangle_{ind} = 0 \quad (29)$$

from which it follows that

$$\sum_{\mathbf{r}} p(\mathbf{r}) = \langle (1 + \xi_p(\mathbf{r})) \rangle_{ind} = 1 , \quad (30)$$

which tells us that  $p(\mathbf{r})$  is properly normalized. Finally, a slightly more involved calculations provides us with a relationship between the parameters of the model and the moments: for  $i \neq j \neq k$ ,

$$\langle \delta r_i \delta r_j \rangle_p = \bar{r}_i (1 - \bar{r}_i) \bar{r}_j (1 - \bar{r}_j) \mathcal{J}_{ij}^p \quad (31a)$$

$$\langle \delta r_i \delta r_j \delta r_k \rangle_p = \bar{r}_i (1 - \bar{r}_i) \bar{r}_j (1 - \bar{r}_j) \bar{r}_k (1 - \bar{r}_k) \mathcal{K}_{ijk}^p. \quad (31b)$$

Virtually identical expressions hold for higher order moments. It is this last set of relationships that make the Sarmanov-Lancaster expansion so useful.

Note that Eqs. (31a) and (31b), along with the expression for the normalized correlation coefficients given in Eq. (13), imply that

$$(1 - \bar{r}_i)(1 - \bar{r}_j) \mathcal{J}_{ij}^p = \rho_{ij}^p \quad (32a)$$

$$(1 - \bar{r}_i)(1 - \bar{r}_j)(1 - \bar{r}_k) \mathcal{K}_{ijk}^p = \rho_{ijk}^p. \quad (32b)$$

These identities will be extremely useful for simplifying expressions later on.

Because the moments are so closely related to the parameters of the distribution, moment matching is especially convenient: to construct a distribution whose moments match those of  $p(\mathbf{r})$  up to some order, one simply needs to ensure that the parameters of that distribution,  $\mathcal{H}_i$ ,  $\mathcal{J}_{ij}$ ,  $\mathcal{K}_{ijk}$ , etc., are identical to those of the true distributions up to the order of interest. In particular, let us write down a new distribution,  $q(\mathbf{r})$ ,

$$q(\mathbf{r}) = p_{ind}(\mathbf{r}) (1 + \xi_q(\mathbf{r})) \quad (33a)$$

$$\xi_q(\mathbf{r}) = \sum_{i < j} \mathcal{J}_{ij}^q \delta r_i \delta r_j + \sum_{i < j < k} \mathcal{K}_{ijk}^q \delta r_i \delta r_j \delta r_k + \dots \quad (33b)$$

We can recover the independent distribution by letting  $\xi_q(\mathbf{r}) = 0$ , and we can recover the pairwise distribution by letting  $\mathcal{J}_{ij}^q = \mathcal{J}_{ij}^p$ . This allows us to compute  $D_{KL}(p||q)$  in the general case, and then either set  $\xi_q$  to zero or set  $\mathcal{J}_{ij}^q$  to  $\mathcal{J}_{ij}^p$ .

Note that expressions analogous to those in Eq. (29-32) exist for averages with respect to  $q(\mathbf{r})$ ; the only difference is that  $p$  is replaced by  $q$ .

### 5.2.1 The KL divergence in the Sarmanov-Lancaster representation

Using Eqs. (25) and (33a) and a small amount of algebra, the KL divergence between  $p(\mathbf{r})$  and  $q(\mathbf{r})$  may be written

$$D_{KL}(p||q) = \frac{1}{\ln 2} \langle f(\xi_p(\mathbf{r}), \xi_q(\mathbf{r})) \rangle_{ind} \quad (34)$$

where

$$f(x, y) \equiv (1+x)[\ln(1+x) - \ln(1+y)] - (x-y). \quad (35)$$

To derive Eq. (34), we used the fact that  $\langle \xi_p \rangle_{ind} = \langle \xi_q \rangle_{ind} = 0$  (see Eq. (29)). The extra term  $(x-y)$  was included to ensure that  $f(x, y)$  and its first derivatives vanish at  $x = y = 0$ , something that greatly simplifies our analysis later on.

Our approach is to Taylor expand the right hand side of Eq. (34) around  $\xi_p = \xi_q = 0$ , compute each term, and then sum the *whole* series. Using  $a_{nm}$  to denote the coefficients of the Taylor series, we have

$$D_{KL}(p||q) = \frac{1}{\ln 2} \sum_{mn} a_{mn} \langle \xi_p(\mathbf{r})^m \xi_q(\mathbf{r})^n \rangle_{ind}. \quad (36)$$

Note that because  $f(x, y)$  and its first derivatives vanish at  $x = y = 0$ , all terms in this sum have  $m + n \geq 2$ .

Because both  $\xi_p$  and  $\xi_q$  are themselves sums, an exact calculation of the terms in Eq. (36) would be difficult. However, as we show in Sec. 5.2.3 (in particular Eqs. (51) and (52)), they can be computed to lowest order in  $N\delta$ , and the result is

$$\begin{aligned} \langle \xi_p(\mathbf{r})^m \xi_q(\mathbf{r})^n \rangle_{ind} &= \frac{1}{\ln 2} \sum_{i < j} \bar{r}_i \bar{r}_j (\rho_{ij}^p)^m (\rho_{ij}^q)^n + \left[ \bar{r}_j (-\bar{r}_i \rho_{ij}^p)^m (-\bar{r}_i \rho_{ij}^q)^n + \bar{r}_i \leftrightarrow \bar{r}_j \right] \\ &+ \frac{1}{\ln 2} \sum_{i < j < k} \bar{r}_i \bar{r}_j \bar{r}_k (\tilde{\rho}_{ijk}^p)^m (\tilde{\rho}_{ijk}^q)^n + \mathcal{O}((N\delta)^4) \end{aligned} \quad (37)$$

where  $\tilde{\rho}_{ijk}^p$  and  $\tilde{\rho}_{ijk}^q$  are given by

$$\tilde{\rho}_{ijk}^x \equiv \rho_{ijk}^x + \rho_{ij}^x + \rho_{ik}^x + \rho_{jk}^x = \frac{\langle r_i r_j r_k \rangle_x - \bar{r}_i \bar{r}_j \bar{r}_k}{\bar{r}_i \bar{r}_j \bar{r}_k}, \quad (38)$$

$x = p, q$ . The last equality in Eq. (38) follows from a small amount of algebra and the definition of the correlation coefficients given in Eq. (13). Equation (37) is valid only when  $m + n \geq 2$ ,

which is the case of interest to us (since the Taylor expansion of  $f(x, y)$  has only terms with  $m + n \geq 2$ ).

The important point about Eq. (37) is that the  $m$  and  $n$  dependence follows that of the original Taylor expansion. Thus, when we insert this equation back into Eq. (36), we recover our original function – all we have to do is interchange the sums. For example, consider inserting the first term in Eq. (37) into Eq. (36),

$$\sum_{m,n} a_{mn} \sum_{i < j} \bar{r}_i \bar{r}_j (\rho_{ij}^p)^m (\rho_{ij}^q)^n = \sum_{i < j} \bar{r}_i \bar{r}_j \sum_{m,n} a_{mn} (\rho_{ij}^p)^m (\rho_{ij}^q)^n = \sum_{i < j} \bar{r}_i \bar{r}_j f(\rho_{ij}^p, \rho_{ij}^q).$$

Performing the same set of manipulations on all of Eq. (37) leads to

$$\begin{aligned} \langle \xi_p(\mathbf{r})^m \xi_q(\mathbf{r})^n \rangle_{ind} &= \frac{1}{\ln 2} \sum_{i < j} \bar{r}_i \bar{r}_j f(\rho_{ij}^p, \rho_{ij}^q) + \bar{r}_j f(-\bar{r}_i \rho_{ij}^p, -\bar{r}_i \rho_{ij}^q) + \bar{r}_i f(-\bar{r}_j \rho_{ij}^p, -\bar{r}_j \rho_{ij}^q) \\ &+ \frac{1}{\ln 2} \sum_{i < j < k} \bar{r}_i \bar{r}_j \bar{r}_k f(\tilde{\rho}_{ijk}^p, \tilde{\rho}_{ijk}^q) + \mathcal{O}(N\delta)^4. \end{aligned} \quad (39)$$

This expression is true in general (except for some technical considerations; see Sec. 5.2.3); to restrict it to the KL divergences of interest we set  $p(\mathbf{r})$  to  $p_{true}(\mathbf{r})$  and  $q(\mathbf{r})$  to either  $p_{ind}(\mathbf{r})$  or  $p_{pair}(\mathbf{r})$ . In the first case,  $q(\mathbf{r}) = p_{ind}(\mathbf{r})$ ,  $\xi_q(\mathbf{r}) = 0$ , which in turn implies that  $\mathcal{J}_{ij}^q = 0$ , and thus  $\rho_{ij}^q = 0$ . Using Eq. (39), we have (to lowest nonvanishing order in  $N\delta$ ),

$$D_{KL}(p_{true}||p_{ind}) = \frac{1}{\ln 2} \sum_{i < j} \bar{r}_i \bar{r}_j f(\rho_{ij}^p, 0) + \mathcal{O}((N\delta)^3). \quad (40)$$

Then, defining

$$g_{ind} \equiv \frac{1}{N(N-1) \ln(2)} \sum_{i < j} \frac{\bar{r}_i}{\delta} \frac{\bar{r}_j}{\delta} f(\rho_{ij}^p, 0), \quad (41)$$

and recalling that  $\delta = \bar{v}\delta t$ , we see that Eq. (40) is equivalent to Eq. (7a).

In the second case,  $q(\mathbf{r}) = p_{pair}(\mathbf{r})$ , the first and second moments of  $p_{pair}(\mathbf{r})$  and  $p_{true}(\mathbf{r})$  are equal. This in turn implies, using Eq. (31), that  $\mathcal{J}_{ij}^q = \mathcal{J}_{ij}^p$ , and thus  $\rho_{ij}^p = \rho_{ij}^q$ . Because  $f(x, x) = 0$  (see Eq. (35)), we see that the first three terms on the right hand side of Eq. (39) – those involving  $i$  and  $j$  but not  $k$  – vanish. The next order term does not vanish, and yields

$$D_{KL}(p_{true}||p_{pair}) = \frac{1}{\ln 2} \sum_{i < j < k} \bar{r}_i \bar{r}_j \bar{r}_k f(\tilde{\rho}_{ijk}^p, \tilde{\rho}_{ijk}^q) + \mathcal{O}((N\delta)^4) . \quad (42)$$

Defining

$$g_{pair} \equiv \frac{1}{N(N-1)(N-2)\ln(2)} \sum_{i < j < k} \frac{r_i}{\delta} \frac{r_j}{\delta} \frac{r_k}{\delta} f(\tilde{\rho}_{ijk}^p, \tilde{\rho}_{ijk}^q) , \quad (43)$$

we see that Eq. (42) reduces to Eq. (7b).

### 5.2.2 Local fields, pairwise couplings and moments

In this section we derive, to leading order in  $N\delta$ , expressions relating the local fields and pairwise couplings of the maximum entropy model,  $h_i$  and  $J_{ij}$ , to the first and second moments. These are the expressions reported in Eq. (15).

To do this, we simply compute the first and second moment under the assumption that  $N\delta$  is small. This calculation proceeds along the same lines as in the previous section, with one extra consideration: the quadratic term in the maximum entropy distribution, Eq. (11), is proportional to  $r_i r_j$ , not  $\delta r_i r_j$ . However, to lowest order in  $N\delta$ , this doesn't matter. That's because

$$\sum_{i < j} J_{ij} r_i r_j = \sum_{i < j} J_{ij} \delta r_i \delta r_j + r_i \sum_{j \neq i} J_{ij} \bar{r}_j + \text{constants} .$$

where  $\bar{r}_i$  is defined as in Eq. (26d) except with  $\mathcal{H}_i^p$  replaced by  $h_i$ , and we used the fact that  $J_{ij} = J_{ji}$ . The second term introduces a correction to the local fields,  $h_i$ . However, the correction is  $\mathcal{O}(N\delta)$ , so we drop it. We should keep in mind, though, that our final expression for  $h_i$  will have corrections of  $\mathcal{O}(N\delta)$ .

Using Eq. (11), but with  $r_i$  replaced by  $\delta r_i$  where it appears with  $J_{ij}$ , we may write (after a small amount of algebra)

$$p_{maxent}(\mathbf{r}) = p_{ind}(\mathbf{r}) \frac{1 + \xi_x(\mathbf{r}) + \psi(\xi_x(\mathbf{r}))}{1 + \langle \xi_x(\mathbf{r}) + \psi(\xi_x(\mathbf{r})) \rangle_{ind}} \quad (44)$$

where  $p_{ind}(\mathbf{r})$  is the same as the function  $p_{ind}(\mathbf{r})$  defined in Eq. (26a), except that  $\mathcal{H}_{ij}^p$  is replaced by  $h$ , the subscript “*ind*” indicates, as usual, an average with respect to  $p_{ind}(\mathbf{r})$ , and the two functions  $\xi_x(\mathbf{r})$  and  $\psi(x)$  are given by

$$\xi_x(\mathbf{r}) \equiv \sum_{i < j} J_{ij} \delta r_i \delta r_j \quad (45)$$

and

$$\psi(x) \equiv e^x - 1 - x. \quad (46)$$

Given this setup, we can use Eqs. (54) and (55) to compute the moments under the maximum entropy model. That's because both  $\psi(x)$  and its first derivative vanish at  $x = 0$ , which are the two conditions required for these equations to be valid. Using also the fact that  $\langle \delta r_i \rangle_{ind} = 0$ , Eqs. (54) and (55) imply that

$$\langle \xi_x(\mathbf{r}) + \psi(\xi_x(\mathbf{r})) \rangle_{ind} = \sum_{i < j} \bar{r}_i \bar{r}_j \psi(J_{ij}) + \mathcal{O}((N\delta)^3) \quad (47a)$$

$$\langle r_i \rangle_{maxent} = (1 + \exp(-h_i))^{-1} + \mathcal{O}(N\delta^2) \quad (47b)$$

$$\langle \delta r_i \delta r_j \rangle_{maxent} = \bar{r}_i \bar{r}_j [\psi(J_{ij}) + J_{ij}] + \mathcal{O}(N\delta^3). \quad (47c)$$

The term “ $+J_{ij}$ ” in Eq. (47c) came from  $\langle \delta r_i \delta r_j \xi_x(\mathbf{r}) \rangle_{ind}$ ; see numerator in Eq. (44). Note that for the second two equations, we used the fact that, to lowest order in  $N\delta$ , the denominator in Eq. (44) is equal to 1.

Finally, using Eq. (16) for the normalized correlation coefficient, dropping the subscript “*maxent*” (since the first and second moments are the same under the maxent and true distributions), and inverting Eqs. (47b) and (47c) to express the local fields and coupling coefficients in terms of the first and second moments, we arrive at Eq (15).

### 5.2.3 Averages of powers of $\xi_p$ and $\xi_q$

Here we compute  $\langle \xi_p^m \xi_q^n \rangle_{ind}$ , which, as can be seen in Eq. (36), is the the key quantity in our perturbation expansion. Our starting point is to (formally) expand the sums that make up  $\xi_p$  and  $\xi_q$  (see Eqs. (26b) and (33b)), which yields

$$\langle \xi_p(\mathbf{r})^m \xi_q(\mathbf{r})^n \rangle_{ind} = \sum_{l=2}^{\infty} \sum_{\{m_1, \dots, m_l\}} \psi_{m_1, \dots, m_l}^{(l)} \sum_{i_1 < \dots < i_l} \langle \delta r_{i_1}^{m_1} \dots \delta r_{i_l}^{m_l} \rangle_{ind}. \quad (48)$$

The sum over  $\{m_1, \dots, m_l\}$  is a sum over all possible configurations of the  $m_i$ . The coefficient  $\psi_{m_1, \dots, m_l}^{(l)}$  are complicated functions of the  $\mathcal{J}_{ij}^p, \mathcal{J}_{ij}^q, \mathcal{K}^{pijk}, \mathcal{K}_{ijk}^q$ , etc. Computing these functions

is straightforward, although somewhat tedious, especially when  $l$  is large; below we compute them only for  $l = 2$  and  $3$ . The reason  $l$  starts at  $2$  is that  $m + n \geq 2$ ; see Eq. (36).

What we will show is that all terms with superscript  $(l)$  are  $\mathcal{O}(\delta^l)$ . To do this, we first note that, because the right hand side of Eq. (48) is an average with respect to the independent distribution, the average of the product is the product of the averages,

$$\langle \delta r_{i_1}^{m_1} \delta r_{i_2}^{m_2} \dots, \delta r_{i_l}^{m_l} \rangle_{ind} = \langle \delta r_{i_1}^{m_1} \rangle_{ind} \langle \delta r_{i_2}^{m_2} \rangle_{ind} \dots, \langle \delta r_{i_l}^{m_l} \rangle_{ind}. \quad (49)$$

Then, using the fact that  $\delta r_i = (1 - \bar{r}_i)$  with probability  $\bar{r}_i$  and  $\delta r_i = (1 - \bar{r}_i)$  with probability  $(1 - \bar{r}_i)$  (see Eq. (26c)), we have

$$\langle \delta r_i^m \rangle_{ind} = \bar{r}_i (1 - \bar{r}_i)^m + (1 - \bar{r}_i) (-\bar{r}_i)^m = \bar{r}_i (1 - \bar{r}_i)^m \left[ 1 - \left( \frac{-\bar{r}_i}{1 - \bar{r}_i} \right)^{m-1} \right]. \quad (50)$$

The significance of this expression is that, for  $m > 1$ ,  $\langle \delta r_i^m \rangle_{ind} \sim \mathcal{O}(r_i) \sim \mathcal{O}(\delta)$ , independent of  $m$ . Consequently, if all the  $m_i$  in Eq. (49) are greater than  $1$ , then the right hand side is  $\mathcal{O}(\delta^l)$ . This shows that, as promised above, the superscript  $(l)$  labels the order of the terms.

As we saw in Sec. 5.2.1, we need to go to third order in  $\delta$ , which means we need to compute the terms on the right hand side of Eq. (48) with  $l = 2$  and  $3$ . Let us start with  $l = 2$ , which picks out only on those terms with two unique indices. Examining the expressions for  $\xi_p$  and  $\xi_q$  given in Eqs. (26b) and (33b), we see that we must keep only terms involving  $\mathcal{J}_{ij}$ , since  $\mathcal{K}_{ijk}$  has three indices, and higher order terms have more. Thus, the  $l = 2$  contribution to the average in Eq. (48), which we denote  $\langle \xi_p(\mathbf{r}) \xi_q(\mathbf{r}) \rangle_{ind}^{(2)}$ , is given by

$$\langle \xi_p(\mathbf{r})^m \xi_q(\mathbf{r})^n \rangle_{ind}^{(2)} = \sum_{i < j} \left\langle \left( \mathcal{J}_{ij}^p \delta r_i \delta r_j \right)^m \left( \mathcal{J}_{ij}^q \delta r_i \delta r_j \right)^n \right\rangle_{ind}.$$

Pulling  $\mathcal{J}_{ij}^p$  and  $\mathcal{J}_{ij}^q$  out of the averages, using Eq. (32a) to eliminate  $\mathcal{J}_{ij}^p$  and  $\mathcal{J}_{ij}^q$  in favor of  $\rho_{ij}^p$  and  $\rho_{ij}^q$ , and applying Eq. (50) (while throwing away some of the terms in that equation that are higher than second order in  $\delta$ ), the above expression may be written

$$\langle \xi_p(\mathbf{r})^m \xi_q(\mathbf{r})^n \rangle_{ind}^{(2)} = \sum_{i < j} \bar{r}_i \bar{r}_j (\rho_{ij}^p)^m (\rho_{ij}^q)^n [1 - (-\bar{r}_i)^{m+n-1} - (-\bar{r}_i)^{m+n-1}]. \quad (51)$$



Note that we were not quite consistent in our ordering with respect to  $\delta$ , in the sense that we kept some higher order terms and not others. We did this so that we could use  $\rho_{ij}$  rather than  $\mathcal{J}_{ij}$ , as the former are directly observable.

For  $l = 3$  the calculation is more involved, but not substantially so. Including terms with exactly three unique indices in the sum on the right hand side of Eq. (48) gives us

$$\begin{aligned} \langle \xi_p(\mathbf{r})^m \xi_q(\mathbf{r})^n \rangle_{ind}^{(3)} = & \sum_{i < j < k} \left\langle \left( \mathcal{K}_{ijk}^p \delta r_i \delta r_j \delta r_k + \mathcal{J}_{ij}^p \delta r_i \delta r_j + \mathcal{J}_{ik}^p \delta r_i \delta r_k + \mathcal{J}_{jk}^p \delta r_j \delta r_k \right)^m \right. \\ & \left. \left( \mathcal{K}_{ijk}^q \delta r_i \delta r_j \delta r_k + \mathcal{J}_{ij}^q \delta r_i \delta r_j + \mathcal{J}_{ik}^q \delta r_i \delta r_k + \mathcal{J}_{jk}^q \delta r_j \delta r_k \right)^n \right\rangle_{ind}. \end{aligned} \quad (52)$$

This expression is not quite correct, since some of its terms contain only two unique indices – these are the terms proportional to  $(\mathcal{J}_{ij}^p)^m (\mathcal{J}_{ij}^q)^n$  – whereas it should contain only terms with exactly three unique indices. Fortunately, this turns out not to matter, for reasons we discuss at the end of the section.

To perform the averages in Eq. (52), we would need to use multinomial expansions, and then average over the resulting powers of  $\delta r$ 's. For the latter, we can work to lowest order in the  $\delta r_i$ , which means we only take the first term in Eq. (50). This amounts to replacing every  $\delta r_i$  with  $1 - \bar{r}_i$  (and similarly for  $j$  and  $k$ ), and in addition multiplying the whole expression by an overall factor of  $\bar{r}_i \bar{r}_j \bar{r}_k$ . For example, if  $m = 1$  and  $n = 2$ , one of the terms in the multinomial expansion is  $\mathcal{K}_{ijk}^p \mathcal{J}_{ij}^q \mathcal{J}_{ik}^q \langle \delta r_i^3 \delta r_j^2 \delta r_k^2 \rangle_{ind}$ . This average would yield, using Eq. (50) and considering only the lowest order term,  $\bar{r}_i \bar{r}_j \bar{r}_k (1 - \bar{r}_i)^3 (1 - \bar{r}_j)^2 (1 - \bar{r}_k)^2$ .

This procedure also is not quite correct, since terms with only one factor of  $\delta r_i$ , which average to zero, are replaced with  $1 - \bar{r}_i$ . This also turns out not to matter; again, we discuss why at the end of the section.

We can, then, go ahead and use the above “replace blindly” algorithm. Note that the factors of  $1 - \bar{r}_i$ ,  $1 - \bar{r}_j$  and  $1 - \bar{r}_k$  turn  $\mathcal{J}_{ij}$  and  $\mathcal{K}_{ijk}$  into normalized correlation coefficients (see Eq. (32)), which considerably simplifies our equations. Using also Eq. (38) for  $\tilde{\rho}_{ijk}$ , Eq. (52) becomes

$$\langle \xi_p(\mathbf{r})^m \xi_q(\mathbf{r})^n \rangle_{ind}^{(3)} = \sum_{i < j < k} \bar{r}_i \bar{r}_j \bar{r}_k (\tilde{\rho}_{ijk}^p)^m (\tilde{\rho}_{ijk}^q)^n. \quad (53)$$

We can now combine Eqs. (51) and (53), and insert them into Eq. (48). This gives us the

first two terms in the perturbative expansion of  $\langle \xi_p(\mathbf{r})^m \xi_q(\mathbf{r})^n \rangle_{ind}$ ; the result is written down in Eq. (37) above.

Why can we ignore the overcounting associated with terms in which an index appears exactly zero or one times? We clearly can't do this in general, because for such terms, replacing  $\delta r_i$  with  $1 - \bar{r}_i$  fails – either because the terms didn't exist in the first place (when one of the indices never appeared) or because they averaged to zero (when an index appeared exactly once). In our case, however, such terms do not appear in the Taylor expansion. To see why, note first of all that, because of the form of  $f(x, y)$ , its Taylor expansion can be written  $(x - y)^2 \tilde{f}(x, y)$  where  $\tilde{f}(x, y)$  is finite at  $x = y$  (see Eq. (35)). Consequently, the expression inside the sum over  $i, j$  and  $k$  in Eq. (52) should really contain a multiplicative factor that arises from  $(\xi_p - \xi_q)^2$ , and thus has the form

$$\left( (\mathcal{K}_{ijk}^p - \mathcal{K}_{ijk}^q) \delta r_i \delta r_j \delta r_k + (\mathcal{J}_{ij}^p - \mathcal{J}_{ij}^q) \delta r_i \delta r_j + (\mathcal{J}_{ik}^p - \mathcal{J}_{ik}^q) \delta r_i \delta r_k + (\mathcal{J}_{jk}^p - \mathcal{J}_{jk}^q) \delta r_j \delta r_k \right)^2.$$

As we saw in the previous section, we are interested in the third order term only to compute  $D_{KL}(p_{true} || p_{pair})$ , for which  $\mathcal{J}_{ij}^p = \mathcal{J}_{ij}^q$ . Therefore, the above multiplicative factor reduces to  $(\mathcal{K}_{ijk}^p - \mathcal{K}_{ijk}^q)^2 (\delta r_i \delta r_j \delta r_k)^2$ . It is that last factor of  $(\delta r_i \delta r_j \delta r_k)^2$  that is important, since it guarantees that every term in the Taylor expansion will have all indices appearing at least twice. Therefore, although Eq. (52) is not true in general, it is valid for our analysis.

We end this section by pointing out that there is a very simple procedure for computing averages to second order in  $\delta$ . Consider a function  $\phi(\xi_p, \xi_q)$  such that  $\phi(\xi_p, \xi_q)$  and its first derivatives vanish at  $\xi_p = \xi_q = 0$ . Then, based on the above analysis, we have

$$\langle \phi(\xi_p, \xi_q) \rangle_{ind} = \sum_{i < j} \bar{r}_i \bar{r}_j \phi(\mathcal{J}_{ij}^p, \mathcal{J}_{ij}^q) + \mathcal{O}((N\delta)^3). \quad (54)$$

Two easy corollaries of this are: for  $k$  and  $l$  positive integers,

$$\langle \delta r_i^k \phi(\xi_p, \xi_q) \rangle_{ind} = \sum_{j \neq i} \bar{r}_i \bar{r}_j \phi(\mathcal{J}_{ij}^p, \mathcal{J}_{ij}^q) + \mathcal{O}(N^2 \delta^3) \quad (55a)$$

$$\langle \delta r_i^k \delta r_j^l \phi(\xi_p, \xi_q) \rangle_{ind} = \bar{r}_i \bar{r}_j \phi(\mathcal{J}_{ij}^p, \mathcal{J}_{ij}^q) + \mathcal{O}(N \delta^3) \quad (55b)$$

where the sum in Eq. (55a) run over  $j$  only, and we used the fact that both  $\mathcal{J}_{ij}^p$  and  $\mathcal{J}_{ij}^q$  are symmetric with respect to the interchange of  $i$  and  $j$ .

### 5.3 Generating synthetic data

As can be seen in Eq. (10), synthetic data depends on three sets of parameters:  $h_i^{true}$ ,  $J_{ij}^{true}$ , and  $K_{ijk}^{true}$ . Here we describe how they were generated.

Our first step was to generate the  $h_i^{true}$ . To do that, we chose a vector  $\mathbf{r}^* = [r_1^* \dots r_{N^*}^*]$  (where, recall,  $N^* = 15$  is the number of neurons in our base distribution), from an exponential distribution with mean 0.02. From this we chose the local field according to Eq. (15a),

$$h_i^{true} = -\log\left(\frac{1}{r_i^*} - 1\right).$$

In the perturbative regime, a distribution generated with these values of the local fields will have firing rates approximately equal to the  $r_i^*$ ; see Eq. (15a) and Fig. 6. We then draw  $J_{ij}^{true}$  and  $K_{ijk}^{true}$  from Gaussian distributions with means equal to 0.05 and 0.02 and standard deviations of 0.8 and 0.5, respectively. Using non-zero values for  $\mathcal{K}_{ijk}$ , means that the distribution is not pairwise.

### 5.4 Bin size and the correlation coefficients

One of our main claims is that  $\Delta_N$  is linear in bin size,  $\delta t$ . This is true, however, only if  $g_{ind}$  and  $g_{pair}$  are independent of  $\delta t$ , as can be seen from Eq. (3). In this section we show that independence is satisfied if  $\delta t$  is smaller than the typical correlation time of the responses. For  $\delta t$  larger than such correlation times,  $g_{ind}$  and  $g_{pair}$  do depend on  $\delta t$ , and  $\Delta_N$  is no longer linear in  $\delta t$ . Note, though, that the correlation time is always finite, so there will always be a bin size below which the linear relationship,  $\Delta \sim \delta t$ , is guaranteed.

Examining Eqs. (41) and (43), we see that  $g_{ind}$  and  $g_{pair}$  depend on the normalized correlation coefficients,  $\rho_{ij}$ ,  $\tilde{\rho}_{ijk}$  (we drop superscripts, since our discussion will be generic). Thus, to understand how  $g_{ind}$  and  $g_{pair}$  depend on bin size, we need to understand how the normalized correlation coefficients depend on bin size.

We start with the second order correlation coefficient, since it is simplest. The corresponding cross-correlogram, which we denote  $C_{ij}(\tau)$ , is given by

$$C_{ij}(\tau) = \frac{1}{\nu_i \nu_j} \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{kl} \delta(t_i^k - t_j^l - \tau) \quad (56)$$

where  $t_i^k$  is the time of the  $k^{\text{th}}$  spike on neuron  $i$  (and similarly for  $t_j^l$ ) and  $\delta(\cdot)$  is the Dirac  $\delta$ -function. The normalization in Eq. (56) is slightly non-standard – more typical is to divide by something with units of firing rate ( $\nu_i$ ,  $\nu_j$  or  $(\nu_i\nu_j)^{1/2}$ ), to give units of spikes/s. The normalization we use is convenient, however, because in the limit of large  $\tau$ ,  $C_{ij}(\tau)$  approaches one.

It is slightly tedious, but otherwise straightforward, to show that when  $\delta t$  is sufficiently small that only one spike can occur in a time bin,  $\rho_{ij}$  is related to  $C_{ij}(\tau)$  via

$$\rho_{ij} = \frac{1}{\delta t} \int_{-\delta t}^{\delta t} d\tau (1 - |\tau|/\delta t) (C_{ij}(\tau) - 1). \quad (57)$$

The (unimportant) factor  $(1 - |\tau|/\delta t)$  comes from the fact that the spikes occur at random locations within a bin.

Equation (57) has a simple interpretation:  $\rho_{ij}$  is the average height of the central peak of the cross-correlogram relative to baseline. How strongly  $\rho_{ij}$  depends on  $\delta t$  is thus determined by the shape of the cross-correlogram. If it is smooth, then  $\rho_{ij}$  approaches a constant as  $\delta t$  becomes small. If, on the other hand, there is a sharp peak at  $\tau = 0$ , then  $\rho_{ij} \sim 1/\bar{\nu}\delta t = 1/\delta$  for small  $\delta t$ , so long as  $\delta t$  is larger than the width of the peak. (The factor of  $\bar{\nu}$  included in the scaling is approximate; it is a placeholder for an effective firing rate that depends on the indices  $i$  and  $j$ . It is, however, sufficiently accurate for our purposes.) A similar relationship exists between the third order correlogram and the correlation coefficient. Thus,  $\tilde{\rho}_{ijk}$  is also independent of  $\delta t$  in the small  $\delta t$  limit, whereas if the central peak is sharp it scales as  $1/\delta^2$ .

The upshot of this analysis is that the shape of the cross-correlogram has a profound effect on the correlation coefficients and, therefore, on  $\Delta_N$ . What is the shape in real networks? The answer typically depends on the physical distance between cells. If two neurons are close, they are likely to receive common input and thus exhibit a narrow central peak in their cross-correlogram. If, on the other hand, the neurons are far apart, they are less likely to receive common input. In this case, the correlations come from external stimuli, so the central peak tends to have a characteristic width given by the temporal correlation time of the stimulus, typically 100s of milliseconds.

Although clearly both kinds of cross-correlograms exist in any single population of neurons, it is convenient to analyze them separately. We have already considered networks in which the cross-correlograms were broad and perfectly flat, so that the correlation coefficients were

strictly independent of bin size. Based on our analysis in Sec. 5.2, we can also consider the opposite extreme: networks in which the cross-correlograms (both second and higher order) among nearby neurons exhibit sharp peaks while those among distant neurons are uniformly equal to 1. In this regime, the correlation coefficients depend on  $\delta t$ : as discussed above, the second order ones scale as  $1/\delta$  and the third as  $1/\delta^2$ . This means that the arguments of  $f(\rho_{ij}, 0)$  and  $f(\tilde{\rho}_{ijk}^{true}, \tilde{\rho}_{ijk}^{pair})$  are large. From the definition of  $f(x, y)$  in Eq. (35), in this regime both are approximately linear in their arguments (ignoring log corrections). Consequently,  $f(\rho_{ij}, 0) \sim 1/\delta$  and  $f(\tilde{\rho}_{ijk}^{true}, \tilde{\rho}_{ijk}^{pair}) \sim 1/\delta^2$ . This implies that  $g_{ind}$  and  $g_{pair}$  scale as  $N\delta$  and  $N^2\delta$ , respectively, and so  $\Delta_N \sim N$ , independent of  $\delta$ . Thus, if the bin size is large compared the correlation time,  $\Delta_N$  will be approximately independent of bin size.

## 5.5 Assessing goodness of fit for independence across time assumption

In this section we derive the expression for  $\gamma$  given in Eq. (22). Our starting point is its definition, Eq. (21). It is convenient to define  $\mathbf{R}$  to be a concatenation of the responses in  $M$  time bins,

$$\mathbf{R} \equiv (\mathbf{r}^1, \mathbf{r}^2, \dots, \mathbf{r}^M) \quad (58)$$

where, as in Sec. 3, the superscript labels time, so  $\mathcal{P}(\mathbf{R})$  is the full, temporally correlated, distribution.

With this definition, we may write the numerator in Eq. (21) as

$$D_{KL}(\mathcal{P}(\mathbf{R}) \parallel \prod_t p_{pair}(\mathbf{r}^t)) = -S_{true}^M - \sum_t \sum_{\mathbf{r}} p_{true}^t(\mathbf{r}) \log_2 p_{pair}(\mathbf{r}) \quad (59)$$

where  $S_{full}^M$  is the entropy of the full distribution,  $\mathcal{P}(\mathbf{R})$ , the last sum follows from a marginalization over all but one element of  $\mathcal{P}(\mathbf{R})$ , and  $p_{true}^t(\mathbf{r})$  is the true distribution at time  $\mathbf{r}$ . Note that  $p_{pair}(\mathbf{r})$  is independent of time, since it is computed from a distribution averaged over all bins. That distribution, which we have called  $p_{true}(\mathbf{r})$ , is given in terms of  $p_{true}^t(\mathbf{r})$  as

$$p_{true}(\mathbf{r}) = \frac{1}{M} \sum_t p_{true}^t(\mathbf{r}).$$

Inserting this definition into Eq. (59) eliminates the sum over  $t$ , and replaces it with  $M p_{true}(\mathbf{r})$ . For simplicity we consider the maximum entropy pairwise model. In this case, because  $p_{pair}(\mathbf{r})$  is in the exponential family, and the first and second moments are the same under the true and

maximum entropy distributions, we can replace  $p_{true}(\mathbf{r})$  with  $p_{maxent}(\mathbf{r})$ . Consequently, Eq. (59) becomes

$$D_{KL}(\mathcal{P}(\mathbf{R})||\prod_t p_{pair}(\mathbf{r}^t)) = MS_{maxent} - S_{true}^M.$$

This gives us the numerator in the expression for  $\gamma$  (Eq. (21)). The denominator,  $D_{KL}(p_{true}||p_{ind})$ , is equal to  $S_{ind} - S_{true}$  (see Eq. (4)). This leads to

$$\gamma = \frac{M(S_{maxent} - S_{true})}{M(S_{ind} - S_{true})} + \frac{MS_{true} - S_{true}^M}{M(S_{ind} - S_{true})}. \quad (60)$$

where we added and subtracted  $MS_{true}$  to the numerator.

The first term on the right hand side of Eq. (60) we recognize, from Eqs. (3), (4) and (5), as  $\Delta_N$ . To cast the second into a reasonable form, we define  $S_{ind}^M$  to be the entropy of the distribution that retains the temporal correlations within each neuron but is independent across neurons. Then, adding and subtracting this quantity to the numerator in Eq. (60), and also adding and subtracting  $MS_{ind}$ , we have

$$\gamma = \Delta_N + \frac{(S_{ind}^M - S_{true}^M) - M(S_{ind} - S_{true}) + (MS_{ind} - S_{ind}^M)}{M(S_{ind} - S_{true})}. \quad (61)$$

The key observation is that if  $M\delta \ll 1$ , then

$$S_{ind}^M - S_{true}^M = g_{ind}N(N-1)(M\delta)^2.$$

Comparing this with Eq. (7a), we see that  $S_{ind}^M - S_{true}^M$  is a factor of  $M^2$  times larger than  $S_{ind} - S_{true}$ . We thus have

$$\gamma = \Delta_N + (M-1) + \frac{MS_{ind} - S_{ind}^M}{M(S_{ind} - S_{true})}. \quad (62)$$

Again assuming  $M\delta \ll 1$ , and defining  $h(x) \equiv -x \log_2 x - (1-x) \log_2 (1-x)$ , the last term in this expression may be written

$$MS_{ind} - S_{ind}^M = M \sum_i h(r_i) - \sum_i h(Mr_i) \approx M \sum_i r_i \log M = N\delta M \log_2 M. \quad (63)$$

Inserting this into Eq. (62) and using Eq. (7a) yields Eq. (22).

We have assumed here that  $M\delta \ll 1$ ; what happens when  $M\delta \sim 1$ , or larger? To answer this, we rewrite Eq. (60) as

$$\gamma = \frac{S_{maxent} - S_{true}^M/M}{\Delta_N}. \quad (64)$$

We argue that in general, as  $M$  increases,  $S_{true}^M/M$  becomes increasingly different from  $S_{maxent}$ , since the former was derived under the assumption that the responses at different time bins were independent. Thus, Eq. (22) should be considered a lower bound on  $\gamma$ .

## References

- [1] Rieke F, Warland D, de Ruyter van Steveninck RR, Bialek W (1997) Spikes: exploring the neural code. Cambridge, MA: MIT press.
- [2] Russ WP, Lowery DM, Mishra P, Yaffe MB, Ranganathan R (2005) Natural like function in artificial ww domains. *Nature* 437:579–583.
- [3] Socolich M, Lockless S, Russ W, Lee H, H GK, et al. (2005) Evolutionary information for specifying a protein fold. *Nature* 437:512–518.
- [4] Oates JF (1987) Food distribution and foraging behavior. In: Smuts B, Cheney DL, Seyfarth RM, Wrangham RW, Struhsaker TT, editors, *Primate societies*, Chicago: University of Chicago Press. pp. 197–209.
- [5] Wrangham RW (1987) Evolution of social structure. In: Smuts B, Cheney DL, Seyfarth RM, Wrangham RW, Struhsaker TT, editors, *Primate societies*, Chicago: University of Chicago Press. pp. 282–298.
- [6] Eisenberg JF, Muckenhirn NA, Rundran R (1972) The relation between ecology a social structure in primates. *Science* 176:863–874.
- [7] Schneidman E, Berry MJn, Segev R, Bialek W (2006) Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature* 440:1007–1012.
- [8] Shlens J, Field G, Gauthier JL, Grivich MI, Petrusca D, et al. (2006) The structure of multi-neuron firing patterns in primate retina. *J Neurosci* 26:8254–8366.
- [9] Tang A, Jackson D, Hobbs J, Chen W, Smith JL, et al. (2008) A maximum entropy model applied to spatial and temporal correlations from cortical networks in vitro. *J Neurosci* 28:505–518.

- [10] Bethge M, Berens P (2008) Near-maximum entropy models for binary neural representations of natural images. In: Platt J, Koller D, Singer Y, Roweis S, editors, *Advances in Neural Information Processing Systems 20*, Cambridge, MA: MIT Press. pp. 97–104.
- [11] Yu S, Huang D, Singer W, Nikolic D (2008) A small world of neuronal synchrony. *Cereb Cortex* :Advance online release, doi:10.1093/cercor/bhn047.
- [12] Kullback S, Leibler RA (1951) On information and sufficiency. *Annals of Mathematical Statistics* 22:79–86.
- [13] Friedman N, Mosenzon O, Slonim N, Tishby N (2001) Multivariate information bottleneck. In: *Proc. of Uncertainty in Artificial Intelligence (UAI-17)*, San Mateo, CA: Morgan Kaufmann Publishers. pp. 152–161.
- [14] Slonim N, Friedman N, Tishby N (2006) Multivariate information bottleneck. *Neural Comp* 18:1739–1789.
- [15] Shannon C, Weaver W (1949) *The mathematical theory of communication*. Urbana, Illinois: University of Illinois Press.
- [16] Cover TM, Thomas JA (1991) *Elements of information theory*. New York, NY: John Wiley & Sons.
- [17] Amari S (2008) Measures of correlations under changing firing rates. *Neural Comp* In press.
- [18] Dill KA (1985) Theory for the folding and stability of globular proteins. *Biochemistry* 24:1501–1509.
- [19] Lockless SW, Ranganathan R (1999) Evolutionarily conserved pathways of energetic connectivity in protein families. *Science* 286:295–299.
- [20] Vargas-Madrado E, Lara-Ochoa F, Jimenez-Montano M (1994) A skewed distribution of amino acids at recognition sites of the hypervariable region of immunoglobulins. *J Mol Evol* 38:100–104.
- [21] Sarmanov OV (1962) Maximum correlation coefficient (nonsymmetric case). In: *Selected Translations in Mathematical Statistics and Probability*, Amer. Math. Soc., volume 2. pp. 207–210.



- [22] Sarmanov OV (1963) Maximum correlation coefficient (nonsymmetric case). In: Selected Translations in Mathematical Statistics and Probability, Amer. Math. Soc., volume 4. pp. 271–275.
- [23] Lancaster HO (1958) The structure of bivariate distributions. Ann Math Statistics 29:719736.
- [24] Lancaster HO (1963) Correlation and complete dependence of random variables. Ann Math Statistics 34:1315–1321.
- [25] Bahadur RR (1961) A representation of the joint distribution of responses to n dichotomous items. In: Solomon H, editor, Studies in Item Analysis and Prediction, Stanford University Press. p. 158168.